Discretization and Preconditioning Algorithms for the Euler and Navier-Stokes Equations on Unstructured Meshes

Tim Barth
NASA Ames R.C.
Moffett Field, California USA

Contents

Chapter 1

Symmetrization of Systems of Conservation Laws

This lecture briefly reviews several related topics associated with the symmetrization of systems of conservation laws and quasi-conservation laws:

1. Basic Entropy Symmetrization Theory

2. Symmetrization and eigenvector scaling

3. Symmetrization of the compressible Navier-Stokes equations

4. Symmetrization of the quasi-conservative form of the MHD equations

There are many motivations, some theoretical, some practical, for recasting conservation law equations into symmetric form. Three motivations are listed below. The first motivation is widely recognized while the remaining two are less often appreciated.

1. Energy Considerations. Consider the compressible Navier-Stokes equations in quasi-linear form with $\mathbf{u}$ the vector of conserved variables, $\mathbf{f}^i$ the flux vectors, and M the viscosity matrix

$$\mathbf{u}_{,t} + \mathbf{f}^i_{,\mathbf{u}}\,\mathbf{u}_{,x_i} = (M_{ij}\mathbf{u}_{,x_j})_{,x_i}. \tag{1.1}$$

   In this form, the inviscid coefficient matrices $\mathbf{f}^i_{,\mathbf{u}}$ are not symmetric and the viscosity matrix M is neither symmetric nor positive semi-definite. This makes energy analysis almost impossible. When recast in symmetric form, the inviscid coefficient matrices are symmetric and the viscosity matrix is symmetric positive semi-definite. The energy analysis associated with Friedrichs systems of this type is well-known.

2. Dimensional consistency. As a representative example, consider the time derivative term from (1.1). The weak variational statement associated with this equation requires the integration of terms such as $-\int \mathbf{w}^T_{,t}\mathbf{u}\ dxdt$. When $\mathbf{w}$ and $\mathbf{u}$ reside in the same space of functions, the inner product quantity $\mathbf{w}^T\mathbf{u}$ is dimensionally inconsistent. Consequently, errors made in a computation would depend fundamentally on how the equations have been non-dimensionalized. When recast in symmetric form, the inner product $\mathbf{w}^T\mathbf{u}$ is dimensionally consistent with units of entropy density per unit volume.

3. Eigenvector scaling. Apart from degenerate scalings, any scaling of eigenvectors satisfies the eigenvalue problem. Unfortunately, numerical discretization techniques sometimes place additional demands on the form of right eigenvectors. As noted by Balsara [Bal94] in his study of high order Godunov methods, several of the schemes he studied that interpolate "characteristic" data (see for example Harten *et al.* [HOEC87]) showed accuracy degradation that depended on the specific scaling of the eigenvectors. In the characteristic interpolation approach, the solution data is projected onto the local right eigenvectors of the flux Jacobian, interpolated between cells, and finally transformed back. The interpolant thus depends on the eigenvector form. Symmetrization provides some additional insight into the scaling of eigenvectors. Let $R(\mathbf{n})$ denote the matrix of right eigenvectors associated with the generalized flux Jacobian in the direction $\mathbf{n}$. Using results from entropy symmetrization theory, Sec. 1.2 describes a scaling of right eigenvectors such that the product $R(\mathbf{n})R^T(\mathbf{n})$ is independent of the vector $\mathbf{n}$. This result is used in Sec. 1.4 which discusses the ideal magnetohydrodynamic (MHD) equations. The right eigenvectors associated with these equations exhibit notoriously poor scaling properties, especially near a triple umbilic point where fast, slow and Alfĕn wave speeds coincide, [BW88]. The entropy symmetrization scaling provides a systematic approach to scaling eigenvectors which is unique in the sense described above.

1.1 A Brief Review of Entropy Symmetrization Theory

Consider a system of m coupled first order differential equations in d space coordinates and time which represents a conservation law process. Let $\mathbf{u}(x,t) : \mathbf{R}^d \times \mathbf{R}^+ \mapsto \mathbf{R}^m$ denote the dependent solution variables and $\mathbf{f}(\mathbf{u}) : \mathbf{R}^m \mapsto \mathbf{R}^{m\times d}$ the flux vector

$$\mathbf{u}_{,t} + \mathbf{f}^i_{,x_i} = 0 \tag{1.2}$$

with implied summation on the index i. Additionally, this system is assumed to possess the following properties:

1. Hyperbolicity. The linear combination

$$\mathbf{f}_{,\mathbf{u}}(\mathbf{n}) = n_i\,\mathbf{f}^i_{,\mathbf{u}}$$

   has m real eigenvalues and a complete set of eigenvalues for all $\mathbf{n} \in \mathbf{R}^d$.

2. Entropy Inequality. Existence of a convex entropy pair $U(\mathbf{u}), F(\mathbf{u}) : \mathbf{R}^m \mapsto \mathbf{R}$ such that in addition to (1.2) the following inequality holds

$$U_{,t} + F^i_{,x_i} \leq 0. \tag{1.3}$$

In the standard symmetrization theory [God61, Moc80, Har83b], one seeks a change of variables $\mathbf{u}(\mathbf{v}) : \mathbf{R}^m \mapsto \mathbf{R}^m$ to Eqn. (1.2) so that when transformed

$$\mathbf{u}_{,\mathbf{v}}\mathbf{v}_{,t} + \mathbf{f}^i_{,\mathbf{v}}\mathbf{v}_{,x_i} = 0 \tag{1.4}$$

the matrix $\mathbf{u}_{,\mathbf{v}}$ is symmetric positive-definite and the matrices $\mathbf{f}^i_{,\mathbf{v}}$ are symmetric. This would be a classical Friedrichs system. Clearly, if functions $\mathcal{U}(\mathbf{v}), \mathcal{F}^i(\mathbf{v}) : \mathbf{R}^m \mapsto \mathbf{R}^1$ can be found such that

$$\mathbf{u} = \mathcal{U}_{,\mathbf{v}}, \quad \mathbf{f}^i = \mathcal{F}^i_{,\mathbf{v}}$$

then the matrices

$$\mathbf{u}_{,\mathbf{v}} = \mathcal{U}_{,\mathbf{v},\mathbf{v}}, \quad \mathbf{f}^i_{\mathbf{v}} = \mathcal{F}^i_{,\mathbf{v},\mathbf{v}}$$

are symmetric. To insure positive-definiteness of $\mathbf{u}_{,\mathbf{v}}$ so that mappings are one-to-one, convexity of $\mathcal{U}(\mathbf{v})$ is imposed. Since $\mathbf{v}$ is not yet known, little progress has been made but introducing the Legendre transform

$$U(\mathbf{u}) = \mathbf{u}^T\mathbf{v} - \mathcal{U}(\mathbf{v})$$

followed by differentiation

$$U_{,\mathbf{u}} = \mathbf{v}^T + \mathbf{u}^T\mathbf{v}_{,\mathbf{u}} - \mathcal{U}_{,\mathbf{v}}\mathbf{v}_{,\mathbf{u}} = \mathbf{v}^T$$

yields an explicit expression for the entropy variables $\mathbf{v}$ in terms of the entropy function $U(\mathbf{u})$. Symmetrization and generalized entropy functions are intimately linked via the following two theorems:

Theorem 1.1.1 Godunov [God61] If a hyperbolic system is symmetrized via change of variables, then there exists a generalized entropy pair for the system.

Theorem 1.1.2 Mock [Moc80] If a hyperbolic system is equipped with a generalized entropy pair U, F^i, then the system is symmetrized under the change of variables $\mathbf{v}^T = U_{,\mathbf{u}}$.

For many physical systems, entropy inequalities of the form (1.3) can be derived by appealing directly to the conservation law system and the second law of thermodynamics. Using this strategy, specific entropy functions for the Navier-Stokes and MHD equations are considered in Secs. 1.3, 1.4 respectively.

1.2 Symmetrization and Eigenvector Scaling

In this section, an important property of right (or left) symmetrizable systems is given. Simplifying upon the previous notation, let $A^0 = \mathbf{u}_{,\mathbf{v}}, A^i = \mathbf{f}^i_{,\mathbf{v}}$ and rewrite (1.4)

$$\underbrace{A^0}_{\text{SPD}}\mathbf{v}_{,t} + \underbrace{A^i A^0}_{\text{Symm}}\mathbf{v}_{,x_i} = 0. \tag{1.5}$$

The following theorem states an important property of the symmetric matrix products A^iA^0 symmetrized via the symmetric positive definite matrix A^0.

Theorem 1.2.1 (Eigenvector Scaling) Let $A \in \mathbf{R}^{n\times n}$ be an arbitrary diagonalizable matrix and S the set of all right symmetrizers:

$$S = \{B \in \mathbf{R}^{n\times n} \mid B \text{ SPD, } AB \text{ symmetric}\}.$$

Further, let $R \in \mathbf{R}^{n\times n}$ denote the right eigenvector matrix which diagonalizes A

$$A = R\Lambda R^{-1}$$

with r distinct eigenvalues, $\Lambda = \mathrm{Diag}(\lambda_1 I_{m_1\times m_1}, \lambda_2 I_{m_2\times m_2}, \ldots, \lambda_r I_{m_r\times m_r})$. Then for each $B \in S$ there exists a symmetric block diagonal matrix $T = \mathrm{Diag}(T_{m_1\times m_1}, T_{m_2\times m_2}, \ldots, T_{m_r\times m_r})$ that block scales columns of R, $\tilde{R} = RT$, such that

$$B = \tilde{R}\tilde{R}^T, \quad A = \tilde{R}\Lambda\tilde{R}^{-1}$$

which imply

$$AB = \tilde{R}\Lambda\tilde{R}^T.$$

Proof: The symmetry of B and $A\,B$ implies that

$$A\,B - B\,A^T = R\Lambda R^{-1}B - BR^{-T}\Lambda R^T = 0$$

or equivalently for $Y \in \mathbf{R}^{n\times n}$

$$\Lambda Y - Y\Lambda = 0, \quad Y = R^{-1}BR^{-T}. \tag{1.6}$$

Partition Y into $r \times r$ blocks, $Y_{m_i\times m_j}$, with block dimensions corresponding to eigenvalue multiplicities. Equation (1.6) then reduces to the following set of decoupled systems:

$$\lambda_i I_{m_i\times m_i} Y_{m_i\times m_j} - \lambda_j Y_{m_i\times m_j} I_{m_j\times m_j} = 0, \quad \forall\, i,j \le r. \tag{1.7}$$

or simply

$$(\lambda_i - \lambda_j) Y_{m_i\times m_j} = 0, \quad \forall\, i,j \le r. \tag{1.8}$$

This implies that Y is of block diagonal form since $Y_{m_i\times m_j} = 0, i \neq j$. From the definition $Y = R^{-1}BR^{-T}$, Y is congruent to B, hence symmetric positive definite (SPD). Given the block diagonal structure of Y, the square root factorization exists globally as well as for each diagonal block $Y_{m_i\times m_i} = Y^{1/2}_{m_i\times m_i} Y^{1/2}_{m_i\times m_i}$. This yields the stated theorem with $T = Y^{1/2}$. ■

This theorem is a variant of the well-known theory developed for the commuting matrix equation

$$AX - XA = 0, \quad A, X \in \mathbf{R}^{n\times n},$$

see for example Gantmacher [Gan59]. Note that the general theory addresses the more general situation for which the matrix A can only be reduced to Jordan canonical form.

Remark 1.2.1 From the scaling theorem 1.2.1, the right eigenvectors associated with each A^i can be scaled so that $A^iA^0 = R^i\Lambda_i(R^i)^T$ which yields a revealing form of the symmetric quasilinear form

$$A^0\mathbf{v}_{,t} + R^i\Lambda_i(R^i)^T\mathbf{v}_{,x_i} = 0$$

with $A^0 = R^1(R^1)^T = \cdots = R^d(R^d)^T$.

1.3 Example: Compressible Navier-Stokes Equations

Consider the ideal compressible Navier-Stokes equations, $x \in \mathbf{R}^d$,

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho\mathbf{V} \\ E \end{pmatrix} + \nabla\cdot\begin{pmatrix} \rho\mathbf{V} \\ \rho\mathbf{V}\mathbf{V} + \mathbf{I}\,p \\ \mathbf{V}\,(E+p) \end{pmatrix} = \nabla\cdot\begin{pmatrix} 0 \\ \tau \\ \tau\mathbf{V} - \mathbf{q} \end{pmatrix} \tag{1.9}$$

where $\mathbf{V} \in \mathbf{R}^d$ is the velocity vector, ρ and p the density and pressure of the fluid, E the specific total energy defined as

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho\mathbf{V}^2 \tag{1.10}$$

and $\boldsymbol{\tau}$ is the viscous stress tensor:

$$\boldsymbol{\tau} = \lambda\left(\frac{\partial V_i}{\partial x_i}\right) + \mu\left(\frac{\partial V_i}{\partial x_j} + \frac{\partial V_j}{\partial x_i}\right). \qquad (1.11)$$

In addition, an ideal gas is assumed $p = \rho R T$ as well as Fourier heat conduction $\mathbf{q} = -\kappa \nabla T$. In these equations λ and μ are diffusion coefficients, κ the coefficient of thermal conductivity, γ the ratio of specific heats, and R the ideal gas law constant.

In 1983, Harten [Har83b] proposed the generalized convex entropy function

$$U(\mathbf{u}) = -\rho g(s), \quad g' > 0, \ \frac{g''}{g'} < \gamma^{-1}$$

for the compressible Euler equations. In this equation, s is the thermodynamic entropy of the fluid. This choice was motivated from the well-known entropy transport inequality for the inviscid Euler equations:

$$s_{,t} + \mathbf{V} \cdot \nabla s \geq 0$$

which generalizes to

$$g(s)_{,t} + \mathbf{V} \cdot \nabla g(s) \geq 0, \quad g' > 0.$$

or after combining with the continuity equations

$$(\rho g(s))_{,t} + \nabla \cdot \rho g(s)\mathbf{V} \geq 0.$$

Thus, it becomes clear that $U = -\rho g(s), F^i = -\rho g(s)V_i$ is an exceptable entropy pair. Hughes, Franca, and Mallet [HFM86] removed the arbitrariness of $g(s)$ by showing that symmetrization of the Navier-Stokes equations with heat conduction places the additional restriction that $g(s)$ be at most affine in s, i.e. $g(s) = c_0 + c_1 s$. A convenient choice is given by $U(s) = \frac{-\rho s}{\gamma-1}$ which yields the following entropy variables:

$$\mathbf{v} = U_{,\mathbf{u}} = \begin{pmatrix} -\frac{s}{\gamma-1} + \frac{\gamma+1}{\gamma-1} - \frac{E}{p} \\ \frac{\rho \mathbf{V}}{p} \\ -\frac{\rho}{p} \end{pmatrix}$$

The change of variable matrix $\mathbf{u}_{,\mathbf{v}}$ takes a particularly simple form:

$$\mathbf{u}_{,\mathbf{v}} = \begin{pmatrix} \rho & \rho\mathbf{V}^T & E \\ \rho\mathbf{V} & \rho\mathbf{V}\mathbf{V} + p\mathbf{I} & \rho H\mathbf{V} \\ E & \rho H\mathbf{V}^T & \rho H^2 - \frac{a^2 p}{\gamma-1} \end{pmatrix}$$

with a the sound speed, $a^2 = \gamma p/\rho$, and H the specific total enthalpy, $H = a^2/(\gamma-1)+\mathbf{V}^2/2$.

Consider the application of the Scaling Theorem 1.2.1 to the inviscid Euler terms appearing in the Navier-Stokes equations

$$A^0 \mathbf{v}_{,t} + A^i A^0 \mathbf{v}_{,x_i} = 0 \qquad (1.12)$$

with $A^i = \mathbf{f}^i_{,\mathbf{u}}$ and $A^0 = \mathbf{u}_{,\mathbf{v}}$. It is sufficient to consider symmetrization of the arbitrary linear combinations of the form

$$A(\mathbf{n}) = n_i A^i, \quad \|\mathbf{n}\| = 1 \qquad (1.13)$$

and scaled right eigenvectors $R(\mathbf{n})$ such that

$$A(\mathbf{n}) = R(\mathbf{n})\Lambda(\mathbf{n})R^{-1}(\mathbf{n}), \quad A^0 = R(\mathbf{n})R^T(\mathbf{n}).$$

From this result, the symmetric coefficient matrices are given by

$$A(\mathbf{n})A^0 = R(\mathbf{n})\Lambda(\mathbf{n})R^T(\mathbf{n}).$$

From a derivational point-of-view, it is advantageous to first compute the eigensystem associated with the system in primitive variables, $\mathbf{w} = (\rho, \mathbf{V}, p)^T$,

$$\mathbf{w}_{,t} + \mathbf{u}_{,\mathbf{w}}^{-1} A^i \, \mathbf{u}_{,\mathbf{w}} \, \mathbf{w}_{,x_i} = 0 \tag{1.14}$$

and then to compute the scaling of the right eigenvectors, $r(\mathbf{n})$, of the primitive variable system. Once the scaled right eigenvectors of the primitive variable system have been computed, scaled right eigenvectors of the conservative system are easily recovered from

$$R(\mathbf{n}) = \mathbf{u}_{,\mathbf{w}} \, r(\mathbf{n})$$

with

$$\mathbf{u}_{,\mathbf{w}} = \begin{pmatrix} 1 & \mathbf{0}^T & 0 \\ \mathbf{V} & \rho\mathbf{I} & \mathbf{0} \\ \frac{1}{2}\mathbf{V}^2 & \rho\mathbf{V}^T & \frac{1}{\gamma-1} \end{pmatrix}. \tag{1.15}$$

1.3.1 Entropy Scaled Eigenvectors for the 3-D Euler Equations

Following the procedure described in Theorem 1.2.1, after some tedious algebraic manipulation, the following scaled right eigenvectors of the primitive variable system have been obtained:

Entropy and Shear Waves: $\lambda_{1,2,3} = \mathbf{V} \cdot \mathbf{n}$

$$r_{1,2,3} = \sqrt{\frac{1}{\gamma\rho}} \begin{pmatrix} \sqrt{\gamma - 1}\, \rho\, \mathbf{n}^T \\ a\, [\mathcal{C}] \\ \mathbf{0}^T \end{pmatrix} \tag{1.16}$$

where $[\mathcal{C}(\mathbf{n})] = n_i \epsilon_{ijk}$ and ϵ_{ijk} is the usual alternation tensor.

Acoustic Waves: $\lambda_{\pm} = \mathbf{V} \cdot \mathbf{n} \pm a$

$$r_{\pm} = \sqrt{\frac{1}{2\gamma\rho}} \begin{pmatrix} \rho \\ \pm a\, \mathbf{n} \\ \rho\, a^2 \end{pmatrix}. \tag{1.17}$$

In Fig. 1.1, the Euclidean norm, $\|R(\mathbf{n})R^T(\mathbf{n})\|_E$, is graphed for $\mathbf{n} = (\cos\theta, \sin\theta)^T, \theta \in [0, 2\pi]$ with constant $(\rho, \mathbf{V}, p)$ using entropy scaled eigenvectors and the "naturally" scaled eigenvectors given in Struijs [Str94]. The entropy scaled eigenvectors produce a constant result since $R(\mathbf{n})R^T(\mathbf{n}) = \mathbf{u}_{,\mathbf{v}}$ which is independent of $\mathbf{n}$. Although the naturally scaled eigenvectors seem to differ only in minor ways from the entropy scaled counterparts, the MHD example given in the next section provides a more convincing argument for proper eigenvector scaling.

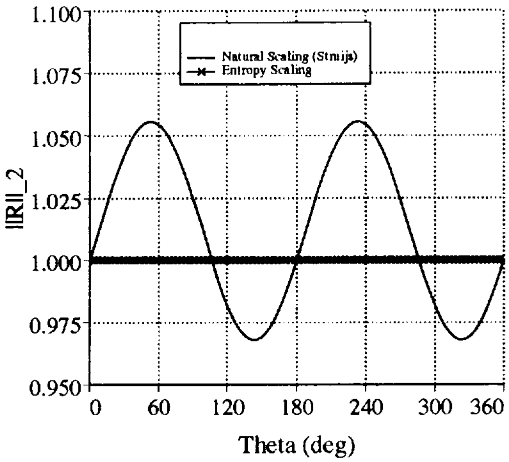

Figure 1.1: Euclidean norm dependency of $R(\mathbf{n})R(\mathbf{n})^T$ on $\mathbf{n}$, $(\rho/rho_\infty = 1, V_1/a_\infty = .8, V_2/a_\infty = -.6, p/p_\infty = 1.4)$

1.4 Example: Quasi-Conservative MHD Equations

As a second example, consider the ideal (nonrelativistic) MHD equations, $x \in \mathbf{R}^d$,

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho\mathbf{V} \\ E \\ \mathbf{B} \end{pmatrix} + \nabla \cdot \begin{pmatrix} \rho\mathbf{V} \\ \rho\mathbf{V}\mathbf{V} + \mathbf{I}\left(p + \frac{1}{2}\mathbf{B}^2\right) - \mathbf{B}\mathbf{B} \\ \mathbf{V}\left(E + p + \frac{1}{2}\mathbf{B}^2\right) - \mathbf{B}\,(\mathbf{V}\cdot\mathbf{B}) \\ \mathbf{V}\mathbf{B} - \mathbf{B}\mathbf{V} \end{pmatrix} = 0. \tag{1.18}$$

In addition to the variables defined for the Navier-Stokes equations, $\mathbf{B} \in \mathbf{R}^d$ is the magnetic field, and E is the specific total energy redefined as

$$E = \frac{p}{\gamma - 1} + \rho\frac{1}{2}\mathbf{V}^2 + \frac{1}{2}\mathbf{B}^2. \tag{1.19}$$

At this point, the equations have been written in conservative (divergence) form. Unlike the Navier-Stokes equations, the MHD equations have the additional time independent constraint $\nabla \cdot \mathbf{B} = 0$. A number of numerical techniques exist for solving problems of this type:

1. Mixed finite element formulations and Lagrange multiplier formulations
2. Finite element penalty formulations.
3. Staggered mesh formulations.
4. Projection formulations.

In the next section, a technique developed by [Pow94] is used to reformulate the MHD equations in quasi-conservative form which obviates the $\nabla \cdot \mathbf{B} = 0$ constraint.

In has been known for some time that negated specific entropy, $-\rho s$, is a convex generalized entropy function for the ideal MHD equations, see Ruggeri [RS81]. Even so, starting from the MHD equations (1.18) in the standard quasi-linear form:

$$\mathbf{u}_{,t} + A^i \mathbf{u}_{,x_i} = 0, \quad A^i = \mathbf{f}^i_{,\mathbf{u}} \tag{1.20}$$

followed by the change of variables to

$$A^0\mathbf{v}_{,t} + A^iA^0\mathbf{v}_{,x_i} = 0, \quad A^0 = \mathbf{u}_{,\mathbf{v}} \tag{1.21}$$

with $\mathbf{v}^T = U_{,\mathbf{u}}$ and $U(\mathbf{u}) = -\frac{\rho s}{\gamma-1}$ does *not* symmetrize the system. For example in 2-D, the following result is obtained:

$$A^1A^0 - (A^1A^0)^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{pB_1^2}{\rho} & -\frac{pB_1}{\rho} & 0 \\ 0 & 0 & 0 & -\frac{pB_1B_2}{\rho} & -\frac{pB_2}{\rho} & 0 \\ 0 & \frac{pB_1^2}{\rho} & \frac{pB_1B_2}{\rho} & 0 & -\frac{pV_2B_2}{\rho} & -\frac{pV_2B_1}{\rho} \\ 0 & \frac{pB_1}{\rho} & \frac{pB_2}{\rho} & \frac{pV_2B_2}{\rho} & 0 & \frac{pV_2}{\rho} \\ 0 & 0 & 0 & -\frac{pV_2B_1}{\rho} & -\frac{pV_2}{\rho} & 0 \end{pmatrix}. \tag{1.22}$$

This failure to symmetrize the MHD system can be explained by the simple observation that the constraint $\nabla\cdot\mathbf{B} = 0$ has not been used. A straightforward derivation of the entropy transport equation for the ideal MHD equations reveals this. For ideal MHD, entropy is given by $s = \log(p\rho^{-\gamma})$ so that

$$ds = -\frac{\gamma}{\rho}d\rho + \frac{1}{p}dp.$$

Inserting terms from (1.18), the following transport equation for smooth flow results:

$$s_{,t} + \mathbf{v}\cdot\nabla s + (\gamma-1)(\mathbf{v}\cdot\mathbf{B})(\nabla\cdot\mathbf{B}) = 0. \tag{1.23}$$

Clearly, the $\nabla\cdot\mathbf{B} = 0$ condition is fundamental in the derivation of the generalized entropy function. The next section considers Powell's method which implicitly satisfies $\nabla\cdot\mathbf{B} = 0$ and permits symmetrization using $U(\mathbf{u}) = -\frac{\rho s}{\gamma-1}$.

1.4.1 Powell's Quasi-Conservative Formulation of the MHD Equations

During an investigation of approximate Riemann solvers for the ideal MHD equations, Powell [Pow94] observed that the coefficient matrices, A^i, appearing in the quasi-linear form are individually rank deficient, i.e. each contains a zero row corresponding to components of the $\mathbf{B}$ field. This degeneracy was removed by replacing the zero eigenvalue and eigenvector with a "divergence wave" eigenvector with velocity eigenvalue similar to the entropy wave. Powell then noted that adding the divergence wave eigenvector was equivalent to writing the MHD equations (1.18) in chain rule form:

$$\begin{aligned}\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{V}) &= 0\\ \frac{\partial\rho\mathbf{V}}{\partial t} + \nabla\cdot(\rho\mathbf{V}\mathbf{V}) + \nabla\left(p + \frac{1}{2}\mathbf{B}^2\right) - \mathbf{B}\cdot\nabla\mathbf{B} - \underline{\mathbf{B}(\nabla\cdot\mathbf{B})} &= 0\\ \frac{\partial E}{\partial t} + \mathbf{V}\cdot\nabla\left(E + p + \frac{1}{2}\mathbf{B}^2\right) + \left(E + p + \frac{1}{2}\mathbf{B}^2\right)\nabla\cdot\mathbf{V} - \mathbf{B}\cdot\nabla(\mathbf{V}\cdot\mathbf{B}) - \underline{(\mathbf{V}\cdot\mathbf{B})(\nabla\cdot\mathbf{B})} &= 0\\ \frac{\partial\mathbf{B}}{\partial t} + \mathbf{V}\cdot\nabla\mathbf{B} + \mathbf{B}\nabla\cdot\mathbf{V} - \mathbf{B}\cdot\nabla\mathbf{V} - \underline{\mathbf{V}(\nabla\cdot\mathbf{B})} &= 0\end{aligned} \tag{1.24}$$

and weakly enforcing $\nabla\cdot\mathbf{B} = 0$ by removing terms proportional to $\nabla\cdot\mathbf{B}$ as underlined in Eqn. 1.24. Powell also noted that this modification changes the nature of the equations in a fundamental way. This is revealed by taking the divergence of the **B** field equations for the original system:

$$\nabla \cdot \left(\frac{\partial \mathbf{B}}{\partial t} + \nabla \cdot (\mathbf{VB} - \mathbf{BV}) \right) = \frac{\partial}{\partial t}(\nabla \cdot \mathbf{B}) = 0 \qquad (1.25)$$

as well as the modified system (1.24) with underlined terms removed:

$$\nabla \cdot \left(\frac{\partial \mathbf{B}}{\partial t} + \mathbf{V} \cdot \nabla \mathbf{B} + \mathbf{B} \nabla \cdot \mathbf{V} - \mathbf{B} \cdot \nabla \mathbf{V} \right) = \frac{\partial}{\partial t}(\nabla \cdot \mathbf{B}) + \nabla \cdot (\mathbf{V} \nabla \cdot \mathbf{B}) = 0. \qquad (1.26)$$

The first form (1.25) states that if $\nabla \cdot \mathbf{B} = 0$ is initially zero then it should remain zero. The second form (1.26) states that $(\nabla \cdot \mathbf{B})/\rho$ is a passive scalar for the system. Any local $\nabla \cdot \mathbf{B}$ is simply advected away. Powell asserts that this is a more numerically more stable process. Finally, note that removal of the underlined terms in (1.24) can be viewed as adding source-like terms proportional to $\nabla \cdot \mathbf{B}$ to the original divergence form (1.18), i.e.

$$\frac{\partial}{\partial t} \begin{pmatrix} \rho \\ \rho \mathbf{V} \\ E \\ \mathbf{B} \end{pmatrix} + \nabla \cdot \begin{pmatrix} \rho \mathbf{V} \\ \rho \mathbf{VV} + \mathbf{I} \left(p + \frac{1}{2}\mathbf{B}^2 \right) - \mathbf{BB} \\ \mathbf{V} \left(E + p + \frac{1}{2}\mathbf{B}^2 \right) - \mathbf{B} \, (\mathbf{V} \cdot \mathbf{B}) \\ \mathbf{VB} - \mathbf{BV} \end{pmatrix} = - \begin{pmatrix} 0 \\ \mathbf{B} \\ \mathbf{V} \cdot \mathbf{B} \\ \mathbf{V} \end{pmatrix} (\nabla \cdot \mathbf{B}). \qquad (1.27)$$

This suggests the quasi-conservative nature of the equations. In the next section, the coefficient matrix form is retained in order to show that the entropy function $U(\mathbf{u}) = -\frac{\rho s}{\gamma - 1}$ does symmetrize Powell's modified MHD equations.

1.4.2 Symmetrization of Powell's Quasi-Conservative MHD Equations

It is again most convenient to write the equations in the primitive variable form $\mathbf{w} = (\rho, \mathbf{V}, p, \mathbf{B})^T$:

$$\mathbf{w}_{,t} + \mathbf{u}_{,\mathbf{w}}^{-1} A_p^i \mathbf{u}_{,\mathbf{w}} \mathbf{w}_{x_i} = 0 \qquad (1.28)$$

where A_p^i denotes a matrix for the modified system. Consider arbitrary combinations, $A_p(\mathbf{n}) = n_i \, A_p^i$ and write Powell's modified coefficient matrices in the following compact form:

$$\mathbf{u}_{,\mathbf{w}}^{-1} A_p(\mathbf{n}) \mathbf{u}_{,\mathbf{w}} = \begin{pmatrix} \mathbf{V} \cdot \mathbf{n} & \rho \mathbf{n}^T & 0 & \mathbf{0}^T \\ \mathbf{0} & (\mathbf{V} \cdot \mathbf{n}) I & \frac{1}{\rho}\mathbf{n} & \frac{1}{\rho}(\mathbf{n}\mathbf{B}^T - \mathbf{B} \cdot \mathbf{n}) \\ 0 & \gamma p \mathbf{n}^T & \mathbf{V} \cdot \mathbf{n} & 0 \\ \mathbf{0} & \mathbf{B}\mathbf{n}^T - \mathbf{B} \cdot \mathbf{n} & \mathbf{0} & (\mathbf{V} \cdot \mathbf{n}) I \end{pmatrix} \qquad (1.29)$$

with

$$\mathbf{u}_{,\mathbf{w}} = \begin{pmatrix} 1 & \mathbf{0}^T & 0 & \mathbf{0}^T \\ \mathbf{V} & \rho \mathbf{I} & \mathbf{0} & \mathbf{0}\mathbf{0}^T \\ \frac{1}{2}\mathbf{V}^2 & \rho \mathbf{V}^T & \frac{1}{\gamma - 1} & \mathbf{B}^T \\ \mathbf{0} & \mathbf{0}\mathbf{0}^T & 0 & I \end{pmatrix}. \qquad (1.30)$$

A straightforward calculation reveals that $U(\mathbf{u}) = -\frac{\rho s}{\gamma - 1}$ *does* symmetrize this system. The entropy variables are given by

$$\mathbf{v} = U_{\mathbf{u}} = \begin{pmatrix} -\frac{s}{\gamma - 1} + \frac{\gamma + 1}{\gamma - 1} - \frac{E}{p} + \frac{\mathbf{B}^2}{2p} \\ \frac{\rho \mathbf{V}}{p} \\ -\frac{\rho}{p} \\ \frac{\rho \mathbf{B}}{p} \end{pmatrix}$$

and the Riemannian matrix $\mathbf{u}_{,\mathbf{v}}$ can be written simply as

$$\mathbf{u}_{,\mathbf{v}} = \begin{pmatrix} \rho & \rho\mathbf{V}^T & E - \frac{1}{2}\mathbf{B}^2 & \mathbf{0}^T \\ \rho\mathbf{V} & \rho\mathbf{V}\mathbf{V} + p\mathbf{I} & \rho H\mathbf{V} & \mathbf{0}\mathbf{0}^T \\ E - \frac{1}{2}\mathbf{B}^2 & \rho H\mathbf{V}^T & \rho H^2 - \frac{a^2 p}{\gamma - 1} + \frac{a^2\mathbf{B}^2}{\gamma} & \frac{p}{\rho}\mathbf{B}^T \\ \mathbf{0} & \mathbf{0}\mathbf{0}^T & \frac{p}{\rho}\mathbf{B} & \frac{p}{\rho}I \end{pmatrix}$$

with a the sound speed, $a^2 = \gamma p/\rho$, and H the specific total enthalpy, $H = a^2/(\gamma-1)+\mathbf{V}^2/2$.

1.4.3 Eigensystem of Powell's Quasi-Conservative MHD Equations

In order to give the eigensystem for the MHD equations, it is useful to define $\mathbf{b} \equiv \mathbf{B}/\sqrt{\rho}$ and the fast and slow speeds:

$$c_{f,s}^2 = \frac{1}{2}\left(a^2 + \mathbf{b}^2\right) \pm \frac{1}{2}\sqrt{(a^2 + \mathbf{b}^2)^2 - 4a^2(\mathbf{b}\cdot\mathbf{n})^2} \tag{1.31}$$

The eigenvalues and eigenvectors are then written compactly as:

Entropy and Divergence Waves: $\lambda_{1,2} = \mathbf{V}\cdot\mathbf{n}$

$$r_1 = \begin{pmatrix} 1 \\ \mathbf{0} \\ 0 \\ \mathbf{0} \end{pmatrix}, \quad r_2 = \begin{pmatrix} 0 \\ \mathbf{0} \\ 0 \\ \mathbf{n} \end{pmatrix} \tag{1.32}$$

Alfvén Waves: $\lambda_{\pm a} = \mathbf{V}\cdot\mathbf{n} \pm (\mathbf{b}\cdot\mathbf{n})$

$$r_{\pm a} = \begin{pmatrix} 0 \\ \pm(\mathbf{n}\times\mathbf{B}) \\ 0 \\ \sqrt{\rho}(\mathbf{n}\times\mathbf{B}) \end{pmatrix} \tag{1.33}$$

Magneto-acoustic Waves: $\lambda_{\pm f,\pm s} = \mathbf{V}\cdot\mathbf{n} \pm c_{f,s}$

$$r_{\pm f,\pm s} = \begin{pmatrix} \rho \\ \pm c_{f,s}\frac{c_{f,s}^2\,\mathbf{n} - (\mathbf{b}\cdot\mathbf{n})\,\mathbf{b}}{c_{f,s}^2 - (\mathbf{b}\cdot\mathbf{n})^2} \\ \rho a^2 \\ c_{f,s}^2\frac{(\mathbf{B}\mathbf{n} - \mathbf{n}\mathbf{B})\,\mathbf{n}}{c_{f,s}^2 - (\mathbf{b}\cdot\mathbf{n})^2} \end{pmatrix} \tag{1.34}$$

In this form, the magneto-acoustic eigenvectors exhibit several forms of degeneracy as carefully described in Roe and Balsara [RB96]. In the next section, the entropy scaled eigenvectors are given. These eigenvectors are similar (but not identical) to the eigenvectors given in Roe and Balsara. The behavior of the new eigenvectors is significantly improved.

1.4.4 Entropy Scaled Eigensystem of Powell's Quasi-Conservative MHD Equations

After consider algebraic manipulation, entropy scaled eigenvectors corresponding to the Powell's quasi-conservative MHD equations have been obtained. Using the notation of Roe and Balsara define

$$\alpha_f^2 = \frac{a^2 - c_s^2}{c_f^2 - c_s^2} \quad \alpha_s^2 = \frac{c_f^2 - a^2}{c_f^2 - c_s^2} \tag{1.35}$$

and $\mathbf{n}^{\perp}$, a unit vector orthogonal to $\mathbf{n}$ lying in the plane spanned by $\mathbf{n}$ and $\mathbf{b}$.

Entropy and Divergence Waves: $\lambda_{1,2} = \mathbf{V} \cdot \mathbf{n}$

$$r_1 = \sqrt{\frac{\gamma - 1}{\gamma}} \begin{pmatrix} \sqrt{\rho} \\ \mathbf{0} \\ 0 \\ \mathbf{0} \end{pmatrix}, \quad r_2 = \sqrt{\frac{1}{\gamma}} \begin{pmatrix} 0 \\ \mathbf{0} \\ 0 \\ a\,\mathbf{n} \end{pmatrix} \tag{1.36}$$

Alfvén Waves: $\lambda_{\pm a} = \mathbf{V} \cdot \mathbf{n} \pm \mathbf{b} \cdot \mathbf{n}$

$$r_{\pm a} = \sqrt{\frac{1}{2}} \begin{pmatrix} 0 \\ \mp \frac{\sqrt{p}}{\rho} \left(\mathbf{n}^{\perp} \times \mathbf{n}\right) \\ 0 \\ \sqrt{\frac{p}{\rho}} \left(\mathbf{n}^{\perp} \times \mathbf{n}\right) \end{pmatrix} \tag{1.37}$$

Fast Magneto-acoustic Waves: $\lambda_{\pm f} = \mathbf{V} \cdot \mathbf{n} \pm c_f$

$$r_{\pm f} = \sqrt{\frac{1}{2\gamma}} \begin{pmatrix} \alpha_f \sqrt{\rho} \\ \pm \frac{\alpha_f\, a^2\, \mathbf{n} + \alpha_s\, a\, ((\mathbf{b} \cdot \mathbf{n}^{\perp})\mathbf{n} - (\mathbf{b} \cdot \mathbf{n})\mathbf{n}^{\perp})}{\sqrt{\rho} c_f} \\ \alpha_f \sqrt{\rho} a^2 \\ \alpha_s a\, \mathbf{n}^{\perp} \end{pmatrix} \tag{1.38}$$

Slow Magneto-acoustic Waves: $\lambda_{\pm s} = \mathbf{V} \cdot \mathbf{n} \pm c_s$

$$r_{\pm s} = \sqrt{\frac{1}{2\gamma}} \begin{pmatrix} \alpha_s \sqrt{\rho} \\ \pm \mathrm{sgn}(\mathbf{b} \cdot \mathbf{n}) \frac{\alpha_s\, a\, (\mathbf{b} \cdot \mathbf{n})\, \mathbf{n} + \alpha_f\, c_f^2\, \mathbf{n}^{\perp}}{\sqrt{\rho} c_f} \\ \alpha_s \sqrt{\rho} a^2 \\ -\alpha_f a\, \mathbf{n}^{\perp} \end{pmatrix} \tag{1.39}$$

Next, the experiment performed in Sec. 1.3.1 is repeated for the MHD equations. In Fig.

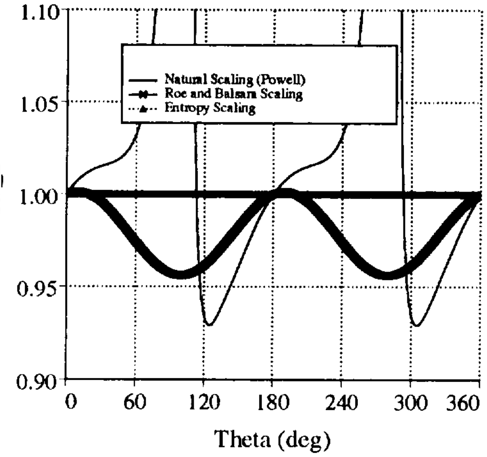

Figure 1.2: Euclidean norm dependency of $R(\mathbf{n})R(\mathbf{n})^T$ on $\mathbf{n}$ for the MHD equations, $(\rho/rho_\infty = 1.3, V_1/a_\infty = .8, V_2/a_\infty = -.6, p/p_\infty = 2.0, B_x = .2, B_y = 1.2)$

1.2, the Euclidean norm, $\|R(\mathbf{n})R^T(\mathbf{n})\|_E$, is graphed for $\mathbf{n} = (\cos\theta, \sin\theta)^T, \theta \in [0, 2\pi]$ with constant $(\rho, \mathbf{V}, p, \mathbf{B})$ using entropy scaled eigenvectors, the naturally scaled eigenvectors given in Sec. 1.4.3, and a slightly generalized form of the eigenvectors given in Roe and Balsara [RB96]. The singularities in the natural scaling are clearly seen. Once again note that the entropy scaled eigenvectors produce a constant result since $R(\mathbf{n})R^T(\mathbf{n}) = \mathbf{u}_{,\mathbf{v}}$ which is independent of $\mathbf{n}$.

In conclusion, using the Scaling Theorem 1.2.1, scaled eigenvectors, $R_p(\mathbf{u})$, have been obtained for Powell's quasi-conservative MHD equations such that

$$A_p(\mathbf{n}) = R_p(\mathbf{n})\Lambda_p(\mathbf{n})R_p^{-1}(\mathbf{n}), \quad A^0 = R_p(\mathbf{n})R_p^T(\mathbf{n})$$

and the symmetric coefficient matrices

$$A_p(\mathbf{n})A^0 = R_p(\mathbf{n})\Lambda(\mathbf{n})R_p^T(\mathbf{n}).$$

In future sections, these results will be exploited in the construction of stabilized numerical discretizations.

Chapter 2

Maximum Principles for Numerical Discretizations on Triangulated Domains

One of the best known tools employed in the study of differential equations is the maximum principle. Any function $f(x)$ which satisfies the inequality $f'' > 0$ on the interval $[a, b]$ attains its maximum value at one of the endpoints of the interval. Solutions of the inequality $f'' > 0$ are said to satisfy a maximum principle. Functions which satisfy a differential inequality in a domain Ω and because of the form of the differential equation achieve a maximum value on the boundary $\partial\Omega$ are said to possess a maximum principle. Recall the maximum principle for Laplace's equation. Let $\Delta u \equiv u_{xx}+u_{yy}$ denote the Laplace operator. If a function u satisfies the strict inequality

$$\Delta u > 0 \tag{2.1}$$

at each point in Ω, then u cannot attain its maximum at any interior point of Ω. The strict inequality can be weakened

$$\Delta u \geq 0 \tag{2.2}$$

so that if a maximum value M is attained in the interior of Ω then the entire function must be a constant with value M. Without any change in the above argument, if u satisfies the inequality

$$\Delta u + c_1 u_x + c_2 u_y > 0$$

in Ω, then u cannot attain its maximum at an interior point.

The second model equation of interest is the nonlinear conservation law equation:

$$u_t + (f(u))_x = 0, \quad \frac{df}{du} = a(u) \tag{2.3}$$

In the simplest setting the initial value problem is considered in which the solution is specified along the x-axis, $u(x, 0) = u_0(x)$ in a periodic or compact supported fashion. The solution can be depicted in the $x - t$ plane by a series of converging and diverging characteristic straight lines. From the solution of (2.3) Lax provides the following observation: *the total increasing and decreasing variations of a differentiable solution between any pairs of characteristics are conserved.*, [Lax73].

$$\mathcal{I}(t + t_0) = I(t_0), \quad \mathcal{I}(t) = \int_{-\infty}^{+\infty} \left| \frac{\partial u(x,t)}{\partial x} \right| dx$$

Moreover in the presence of entropy satisfying discontinuities the total variation decreases (information is destroyed) in time.

$$\mathcal{I}(t+t_0) \le I(t_0) \tag{2.4}$$

An equally important consequence of Lax's observation comes from considering a monotonic solution between two non-intersecting characteristics: ***between pairs of characteristics, monotonic solutions remain monotonic***, no new extrema are created. Also from (2.4) it follows that

1. Local maxima are nonincreasing
2. Local minima are nondecreasing

These properties of the differential equations serve as basic design principles for numerical schemes which approximate them. The next section reviews the theory surrounding discrete matrix operators equipped with maximum principles.

2.1 Discrete Maximum Principles for Elliptic Equations

2.1.1 Laplace's Equation on Structured Meshes

Consider Laplace's equation with Dirichlet data

$$\begin{gathered}\mathcal{L}u = 0, x, y \in \Omega \\ u = g, x, y \in \partial\Omega \\ \mathcal{L} = \Delta\end{gathered} \tag{2.5}$$

From the maximum principle property we have that

$$\sup_{x\in\Omega} |u(x,y)| \le \sup_{x\in\partial\Omega} |u(x,y)|$$

For simplicity consider the unit square domain

$$\Omega = \{(x,y) \in R^2 : 0 \le x, y \le 1\}$$

with spatial grid $x_j = j\Delta x$, $y_k = k\Delta x$, and $J\Delta x = 1$. Let $U_{j,k}$ denote the numerical approximation to $u(x_j, y_k)$. It is well known that the standard second order accurate approximation

$$\mathcal{L}_\Delta U = \frac{1}{\Delta x^2}[U_{j+1,k} + U_{j-1,k} + U_{j,k+1} + U_{j,k-1} - 4U_{j,k}] \tag{2.6}$$

exhibits a discrete maximum principle. To see this simply solve for the value at (j,k)

$$U_{j,k} = \frac{1}{4}[U_{j+1,k} + U_{j-1,k} + U_{j,k+1} + U_{j,k-1}].$$

If $U_{j,k}$ achieves a maximum value M in the interior then

$$M = \frac{1}{4}[U_{j+1,k} + U_{j-1,k} + U_{j,k+1} + U_{j,k-1}]$$

which implies that

$$M = U_{j+1,k} = U_{j-1,k} = U_{j,k+1} = U_{j,k-1}$$

Repeated application of this argument for the four neighboring points yields a discrete maximum principle.

2.1.2 Monotone Matrices

The discrete Laplacian operator $-\mathcal{L}_\Delta$ obtained from (2.6) is one example of a *monotone* matrix.

Definition: A matrix $\mathcal{M}$ is a monotone matrix if and only if $\mathcal{M}^{-1} \geq 0$ (all entries are nonnegative).

Theorem 2.1.1 (Monotone Matrix) A sufficient but not necessary condition for $\mathcal{M}$ monotone is that $\mathcal{M}$ be an M-matrix. M-matrices have the sign pattern $\mathcal{M}_{ii} > 0$ for each i, $\mathcal{M}_{ij} \leq 0$ whenever $i \neq j$. In addition $\mathcal{M}$ must either be strictly diagonally dominant

$$\mathcal{M}_{ii} > \sum_{j=1,j\neq i}^{n} |\mathcal{M}_{ij}|, \quad i = 1, 2, ..., n \qquad \text{(strict diagonal dominance)}$$

or else $\mathcal{M}$ must be irreducible and

$$\mathcal{M}_{ii} \geq \sum_{j=1,j\neq i}^{n} |\mathcal{M}_{ij}|, \quad i = 1, 2, ..., n \qquad \text{(diagonal dominance)}$$

with strict inequality for at least one i.

Proof: The proof for strictly diagonally dominant M is straightforward. Rewrite the matrix operator in the following form

$$\begin{aligned} \mathcal{M} &= D - N, \qquad D > 0, N \geq 0 \\ &= [I - ND^{-1}]D \qquad D^{-1} > 0 \\ &= [I - P]D \qquad P \geq 0 \end{aligned} \tag{2.7}$$

From the strict diagonal dominance of $\mathcal{M}$, eigenvalues of $P = ND^{-1}$ are less than unity. This implies that the Neumann series for $[I - P]^{-1}$ is convergent. This yields the desired result:

$$\mathcal{M}^{-1} = D^{-1}[I + P + P^2 + P^3 + ...] \geq 0 \tag{2.8}$$

When $\mathcal{M}$ is not strictly diagonally dominant then $\mathcal{M}$ must be irreducible so that no permutation $\mathcal{P}$ exists such that

$$\mathcal{P}^T \mathcal{M} \mathcal{P} = \begin{bmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} \\ 0 & \mathcal{M}_{22} \end{bmatrix} \quad \text{(reducibility).}$$

This insures that eigenvalues of P are less than unity. Once again, the Neumann series is convergent and the final result follows immediately. ■

2.1.3 Laplace's Equation on Unstructured Meshes

Consider solving the Laplace equation problem (2.5) on a planar triangulation using a Galerkin finite element approximation with linear elemental shape functions. (Results using a finite volume method are identical but are not considered here.) Next pose (2.5) in variational form. Let $\mathcal{S}^h \in H^1$ be the finite-dimensional space of trial functions with bounded energy which satisfy the Dirichlet boundary condition on Γ. Similarly, let $\mathcal{V}^h \in H^1_0$ denote the finite-dimensional space of functions satisfying homogeneous boundary conditions. Find $u \in \mathcal{S}^h$ such that for all $w \in \mathcal{V}^h$

$$\int_\Omega (\nabla u \cdot \nabla w)\, dx = 0 \tag{2.9}$$

with

$$u(x) = g(x), \quad x \in \Gamma.$$

From this simple equation, we have the following remarkable theorem:

Lemma 2.1.1 The C^0 linear Galerkin finite element discretization of the 2-D Laplace equation (2.9) is a monotone discretization if and only if the triangulation is a Delaunay triangulation.

Proof: Consider a single arbitrary simplex $T = \text{simplex}(x_1, x_2, x_3)$ and the discretization of (2.9) in terms of local linear shape functions $N_i(x)$ satisfying $N_i(x_j) = \delta_{ij}$. Using these shape functions $u(x) = \sum_{j=1}^3 N_j(x)u_j,\ x \in T$ and $w(x) = \sum_{j=1}^3 N_j(x)w_j,\ x \in T$. Inserting these expressions into (2.9) yields

$$\int_T \nabla w \cdot \nabla u \; dx = \sum_{i=1}^{3}\sum_{j=1}^{3} w_i \; u_j \; (\nabla N_i \cdot \nabla N_j) \text{ meas}(T). \tag{2.10}$$

These expressions can be collected pairwise for edges surrounding a vertex. After some straightforward manipulation, the following global discretization formula is obtained

$$\int_\Omega (\nabla w \cdot \nabla u) \; dx = \sum_{i=1}^{|V|} w_i \sum_{j \in \mathcal{N}_i} W_{ij} \; (u_i - u_j) = 0 \tag{2.11}$$

where $\mathcal{N}_i$ denotes the set of vertices adjacent to vertex v_i with weights

$$\begin{aligned} W_{ij} &= (\nabla N_i \cdot \nabla N_j)\text{meas}(T) + (\nabla N_i' \cdot \nabla N_j')\text{meas}(T') \\ &= \frac{1}{2}\left(\text{cotan}(\alpha_{ij}) + \text{cotan}(\alpha_{ij}')\right). \end{aligned} \tag{2.12}$$

In this formula, α_{ij} and α_{ij}' are the two angles subtending the edge $e(v_i, v_j)$, see Fig. 2.1.

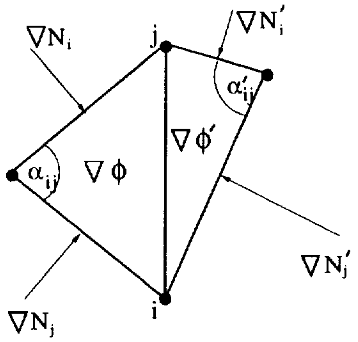

Figure 2.1: Discretization weight geometry for the edge $e(v_i, v_j)$

Since the discretization formula must hold for arbitrary values of w_i at interior vertices, it can be concluded that for all interior vertices v_i

$$\sum_{j \in \mathcal{N}_i} W_{ij} \; (u_i - u_j) = 0. \tag{2.13}$$

Written in this form, the discretization is monotone if all weights are nonnegative, $W_{ij} \geq 0$ and Dirichlet boundary conditions are enforced. Further simplification of the edge weight formula is possible

$$\begin{aligned} W_{ij} &= \frac{1}{2}\left(\mathrm{cotan}(\alpha_{ij}) + \mathrm{cotan}(\alpha'_{ij})\right) \\ &= \frac{1}{2}\left(\frac{\cos(\alpha_{ij})}{\sin(\alpha_{ij})} + \frac{\cos(\alpha'_{ij})}{\sin(\alpha'_{ij})}\right) \\ &= \frac{1}{2}\left(\frac{\sin(\alpha_{ij} + \alpha'_{ij})}{\sin(\alpha_{ij})\sin(\alpha'_{ij})}\right). \end{aligned} \tag{2.14}$$

Since $\alpha_{ij} < \pi$, $\alpha'_{ij} < \pi$, the denominator is always positive, hence nonnegativity requires that $\alpha_{ij} + \alpha'_{ij} \leq \pi$.

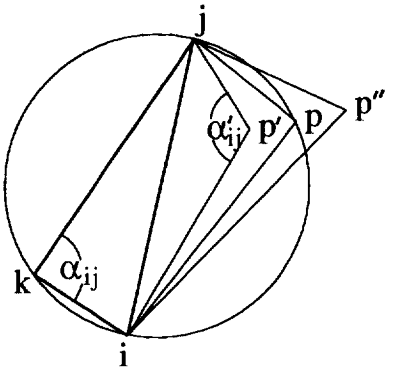

Figure 2.2: Circumcircle test for adjacent triangles, p' interior, p'' exterior, p cocircular.

Some trigonometry reveals that for the configuration shown in Fig. 2.2 with circumcircle passing through $\{v_i, v_j, v_k\}$ the sum $\alpha_{ij} + \alpha'_{ij}$ depends on the location of p with respect to the circumcircle in the following way:

$$\begin{aligned} \alpha_{ij} + \alpha'_{ij} &< \pi, && p \text{ exterior} \\ \alpha_{ij} + \alpha'_{ij} &> \pi, && p \text{ interior} \\ \alpha_{ij} + \alpha'_{ij} &= \pi, && p \text{ cocircular} \end{aligned} \tag{2.15}$$

Also note that by considering the circumcircle passing through $\{v_i, v_j, v_p\}$, similar results would be obtained for v_k. The condition of nonnegativity implies a circumcircle condition for all pairs of adjacent triangles whereby the circumcircle passing through either triangle cannot contain the fourth point. This is precisely the *unique* characterization of the Delaunay triangulation [Del34] which completes the proof. ■

Observe that from equation (2.12) that $\mathrm{cotan}(\alpha) \geq 0$ if $\alpha \leq \pi/2$. Therefore a sufficient but not necessary condition for nonnegativity of the Laplacian weights is that all angles of the mesh be less than or equal to $\pi/2$. This is a standard result in finite element theory [CR73] and applies in two or more space dimensions.

Given Lemma 2.1.1, it becomes straightforward to obtain a discrete maximum principle for Laplace's equation on Delaunay triangulations using a Galerkin finite element approximation.

Theorem 2.1.2 The discrete Laplacian operator obtained from the Galerkin finite element discretization of (2.9) with C^0 linear elements exhibits a discrete maximum principle for arbitrary point sets in two space dimensions if the triangulation of these points is a Delaunay triangulation.

Proof: From Lemma 2.1.1, a one-to-one correspondence exists between nonnegativity of weights and Delaunay triangulation. Assume a Delaunay triangulation of the point set so that for some arbitrary interior vertex v_0 and all adjacent neighbors v_i we have $W_{0i} \geq 0$. Next solve for u_0:

$$u_0 = \frac{\sum_{i \in \mathcal{N}_0} W_{0i} u_i}{\sum_{i \in \mathcal{N}_0} W_{0i}} = \sum_{i \in \mathcal{N}_0} \sigma_i u_i$$

with

$$\sigma_i = \frac{W_{0i}}{\sum_{i \in \mathcal{N}_0} W_{0i}}$$

which satisfies $\sigma_i \geq 0$ and $\sum_{i \in \mathcal{N}_0} \sigma_i = 1$. This implies u_0 is a convex combination of the neighboring values, hence

$$\min_{i \in \mathcal{N}_0} u_i \leq u_0 \leq \max_{i \in \mathcal{N}_0} u_i \tag{2.16}$$

If u_0 attains a maximum value M then all $u_i = M$. Repeated application of (2.16) to neighboring vertices in the triangulation establishes the discrete maximum principle. ■

One can ask if the result concerning Delaunay triangulation and the maximum principle extends to three space dimensions. Unfortunately, the answer is no. This can be demonstrated by counterexample. The resulting formula for the three-dimensional Laplacian edge weight is

$$W_{0i} = \frac{1}{6} \sum_{k=1}^{d(v_0,v_i)} |\Delta \mathbf{R}_{k+\frac{1}{2}}| \operatorname{cotan}(\alpha_{k+\frac{1}{2}}). \tag{2.17}$$

In this formula, $\mathcal{N}_0$ is the set of indices of all adjacent neighbors of v_0 connected by incident edges, k a local cyclic index describing the associated vertices which form a polygon of degree $d(v_0, v_i)$ surrounding the edge $e(v_0, v_i)$, $\alpha_{k+\frac{1}{2}}$ is the face angle between the two faces associated with $\vec{\mathbf{S}}_{k+\frac{1}{2}}$ and $\vec{\mathbf{S}}'_{k+\frac{1}{2}}$ which share the edge $e(v_k, v_{k+1})$ and $|\Delta \mathbf{R}_{k+\frac{1}{2}}|$ is the magnitude of the edge, see Fig. 2.3. A maximum principle is guaranteed if all $W_{0i} \geq 0$. We now will proceed to describe a valid Delaunay triangulation with one or more $W_{0i} < 0$. It

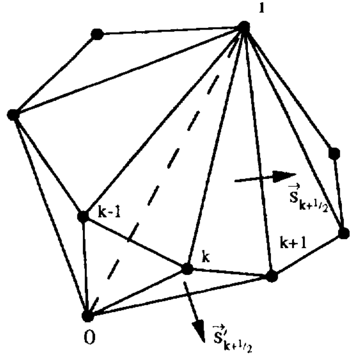

Figure 2.3: Set of tetrahedra sharing interior edge $e(v_0, v_i)$ with local cyclic index k.

will suffice to consider the Delaunay triangulation of N points in which a single point v_0 lies interior to the triangulation and the remaining $N-1$ points describe vertices of boundary faces which completely cover the convex hull of the point set.

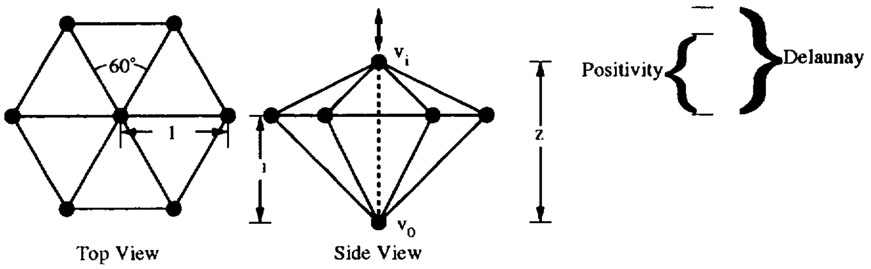

Figure 2.4: Subset of 3-D Delaunay Triangulation that fails to maintain nonnegativity.

Consider a subset of the N vertices, in particular consider an interior edge incident to v_0 connecting to v_i as shown in Fig. 2.4 by the dashed line segment and all neighbors adjacent to v_i on the hull of the point set. In this experiment the height of the interior edge, z, serves as a free parameter. Although it will not be proven here, the remaining $N-8$ points can be placed without conflicting with any of the conclusions obtained for looking at the subset.

It is known that a necessary and sufficient condition for the 3-D Delaunay triangulation is that the circumsphere passing through the vertices of any tetrahedron must be point free [Law86]; that is to say that no other point of the triangulation can lie interior to this sphere. Furthermore a property of locality exists so that only adjacent tetrahedra need be inspected for the satisfaction of the circumsphere test. For the configuration of points shown in Fig. 2.4, convexity of the point cloud constrains $z \geq 1$ and the satisfaction of the circumsphere test requires that $z \leq 2$.

$$1 \leq z \leq 2 \quad \text{(Delaunay Triangulation)}$$

From (2.17), $W_{0i} \geq 0$ if and only if $z < 7/4$.

$$1 \leq z \leq \frac{7}{4}, \quad \text{(Nonnegativity)}$$

This indicates that for $7/4 < z \leq 2$ a valid Delaunay triangulation exists which does not satisfy a discrete maximum principle. In fact, the Delaunay triangulation of 400 points randomly distributed in the unit cube revealed that approximately 25% of the interior edge weights were of the wrong sign (negative).

Keep in mind that from (2.17) the sufficient but not necessary condition for nonnegativity that all face angles be less than or equal to $\pi/2$. This is consistent with the known result from [CR73].

2.2 Discrete Total Variation and Maximum Principles for Hyperbolic Equations

This section examines discrete total variation and maximum principles for scalar conservation law equations. Begin by considering the nonlinear conservation law equation:

$$u_t + (f(u))_x = 0, \; u(x,0) = u_0(x), \quad x,t \in \mathbf{R} \times \mathbf{R}^+$$

which is discretized in the conservation form:

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}(h_{j+\frac{1}{2}} - h_{j-\frac{1}{2}}) \\ &= H(U_{j-l}^n, U_{j-l+1}^n, ..., U_{j+l}) \end{aligned} \tag{2.18}$$

where $h_{j+\frac{1}{2}} = h(U_{j-l+1}, ..., U_{j+l})$ is the numerical flux function satisfying the consistency condition

$$h(U, U, ..., U) = f(U).$$

A finite-difference scheme (2.18) is said to be *monotone* in the sense of Harten, Hyman, and Lax [HHL76] if H is a monotone increasing function of each of its arguments.

$$\frac{\partial H}{\partial U_i}(U_{-k}, ..., U_k) \geq 0 \ \ \forall \ -k \leq i \leq k \qquad \text{(HHL monotonicity)}$$

This is a strong definition of monotonicity. In Crandall and Majda [CM80] it is proven that schemes on Cartesian grids satisfying this condition converge to the physically relevant, entropy satisfying solution. Kröner *el al.* [KRW96] has recently proven a similar result for monotone upwind finite volume schemes on triangulated domains. Unfortunately, HHL monotone schemes in conservation form are at most first order spatially accurate. Very few results are known concerning the convergence of high order accurate approximations. Johnson and Szepessy [JS90] have shown convergence to entropy solutions using streamline diffusion with specialized shock capturing operators. Kröner *et al.* [KSR95] have recently obtained measure-valued convergence of higher order upwind finite volume schemes for scalar conservation laws in several space dimensions.

To circumvent the first order accuracy of monotone schemes, Harten introduced a weaker concept of monotonicity. A grid function U is called monotone if for all i

$$\min(U_{i-1}, U_{i+1}) \leq U_i \leq \max(U_{i-1}, U_{i+1}).$$

A scheme is called monotonicity preserving if monotonicity of U^{n+1} follows from monotonicity of U^n. Observe the close relationship between monotonicity preservation in time and the discrete maximum principle for Laplace's equation (2.16) in space. It follows immediately from the definition of monotonicity preservation that

1. Local maxima are nonincreasing

2. Local minima are nondecreasing

which is a property of the conservation law equation. Using this weaker form of monotonicity Harten [Har83a] introduced the notion of total variation diminishing schemes. Define the total variation in one dimension:

$$TV(U) = \sum_{-\infty}^{\infty} |U_i - U_{i-1}|.$$

A scheme is said to be total variation diminishing (TVD) if

$$TV(U^{n+1}) \leq TV(U^n)$$

This is a discrete analog of the total variation statement (2.4) given for the conservation law equation. Harten has proven that schemes which are HHL monotone are TVD and schemes that are TVD are monotonicity preserving. Furthermore, it can be shown that all *linear* monotonicity preserving schemes are at most first order accurate. Thus high order accurate TVD schemes must necessarily be nonlinear in a fundamental way.

To understand the basic design principles for TVD schemes, assume a one-dimensional periodic grid together with the following numerical scheme in abstract matrix operator form

$$[I + \theta\Delta t\ \widetilde{M}\ D]U^{n+1} = [I - (1-\theta)\Delta t\ M\ D]U^{n} \tag{2.19}$$

where $\widetilde{M}$ and M are matrices which can be nonlinear functions of the solution U. The matrix D denotes the difference operator

$$D\,U = [I - E^{-1}]U = \begin{pmatrix} U_1 - U_J \\ U_2 - U_1 \\ U_3 - U_2 \\ \vdots \\ U_J - U_{J-1} \end{pmatrix}.$$

The scheme (2.19) represents a general family of explicit ($\theta = 0$) and implicit ($\theta = 1$) schemes with arbitrary support. More importantly, schemes written in *conservative* form can be rewritten in this form using (exact) mean value constructions. Using this notation an equivalent definition of the total variation in terms of the L_1 norm is produced

$$TV(U) = \|D\,U\|_1.$$

To analyze the scheme (2.19), multiply by D from the left and regroup.

$$[I + \theta\Delta t\ D\,\widetilde{M}]D\,U^{n+1} = [I - (1-\theta)\Delta t\ D\,M]D\,U^{n}$$

or in symbolic form

$$\mathcal{L}D\,U^{n+1} = \mathcal{R}D\,U^{n} \quad D\,U^{n+1} = \mathcal{L}^{-1}\mathcal{R}D\,U^{n}$$

where invertibility of $\mathcal{L}$ has been assumed. This invertibility will be guaranteed from the diagonal dominance required below. Next take the L_1 norm of both sides and apply matrix-vector inequalities.

$$\begin{aligned} TV(U^{n+1}) = \|D\,U^{n+1}\|_1 &\le \|\mathcal{L}^{-1}\mathcal{R}\|_1\|D\,U^{n}\|_1 \\ &= \|\mathcal{L}^{-1}\mathcal{R}\|_1 TV(U^{n}) \end{aligned} \tag{2.20}$$

This reveals the sufficient condition that $\|\mathcal{L}^{-1}\mathcal{R}\|_1 \le 1$. Recall that the L_1 norm of a matrix is obtained by summing the absolute value of elements of columns of the matrix and choosing the column whose sum is largest. Furthermore, we have the usual matrix inequality $\|\mathcal{L}^{-1}\mathcal{R}\|_1 \le \|\mathcal{L}^{-1}\|_1\|\mathcal{R}\|_1$ so that sufficient TVD conditions are that $\|\mathcal{L}^{-1}\|_1 \le 1$ and $\|\mathcal{R}\|_1 \le 1$. These simple inequalities are enough to recover the TVD criteria of previous investigators, see [Har83b, JL86].

Theorem 2.2.1 A sufficient condition for the scheme (2.19) to be TVD with $\theta = 0$ is that $\mathcal{R}$ be bounded with $\mathcal{R} \ge 0$ (all elements are nonnegative).

Proof: Consider the explicit operator $\mathcal{R} = I - \Delta t\ D M$ and multiply it from the left by the summation vector $s^T = [1, 1, ..., 1]$. It is clear that $s^T D = 0$ so that $s^T \mathcal{R} = s^T$ (columns sum to unity). Because the L_1 norm of $\mathcal{R}$ is the maximum of the sum of absolute values of elements in columns of $\mathcal{R}$, the stated theorem is proven. ■

Next consider the implicit scheme with $\theta = 1$ and sufficient conditions for $\|\mathcal{L}^{-1}\|_1 \leq 1$. From the previous development, one way to do this would be to show that $\mathcal{L}$ is monotone, i.e. $\mathcal{L}^{-1} \geq 0$ with columns that sum to unity.

Theorem 2.2.2 A sufficient condition for the scheme (2.19) to be TVD with $\theta = 1$ is that $\mathcal{L}$ be an M-type monotone matrix, i.e. diagonally dominant with positive diagonal entries and negative off-diagonal entries.

Proof: Consider the implicit operator $\mathcal{L} = I + \Delta t\ D \widetilde{M}$ and multiply it from the left by the summation vector. Again note that $s^T D = 0$ so that

$$s^T \mathcal{L} = s^T \to s^T = s^T \mathcal{L}^{-1}$$

which implies that columns of $\mathcal{L}^{-1}$ sum to unity. The final result is obtained by appealing to Theorem 2.1.1 concerning M-type monotone matrix operators. ■

This general theory reproduces some well known results by Harten [Hart83]. Consider the following explicit scheme in Harten's notation:

$$U_j^{n+1} = U_j^n + C^+_{j+1/2} \Delta_{j+1/2} U^n - C^-_{j-1/2} \Delta_{j-1/2} U^n$$

where $\Delta_{j+1/2} U = U_{j+1} - U_j$. The operator $\mathcal{R}$ in this case has the following banded structure

$$\begin{pmatrix} \ddots & \ddots & 0 & 0 & \ddots \\ \ddots & \ddots & C^+_{j+1/2} & 0 & 0 \\ 0 & C^-_{j-1/2} & 1 - C^+_{j+1/2} - C^-_{j+1/2} & C^+_{j+3/2} & 0 \\ 0 & 0 & C^-_{j+1/2} & \ddots & \ddots \\ \ddots & 0 & 0 & \ddots & \ddots \end{pmatrix}$$

We need only require that this matrix be nonnegative to arrive at Harten's criteria:

$$C^+_{j+1/2} \geq 0$$

$$C^-_{j+1/2} \geq 0$$

$$1 - C^+_{j+1/2} - C^-_{j+1/2} \geq 0$$

Next consider Harten's implicit form:

$$U_j^{n+1} + D^+_{j+1/2} \Delta_{j+1/2} U^{n+1} - D^-_{j-1/2} \Delta_{j-1/2} U^{n+1} = U_j^n$$

In this case $\mathcal{L}$ has the general structure

$$\begin{pmatrix} \ddots & \ddots & 0 & 0 & \ddots \\ \ddots & \ddots & -D^+_{j+1/2} & 0 & 0 \\ 0 & -D^-_{j-1/2} & 1 + D^+_{j+1/2} + D^-_{j+1/2} & -D^+_{j+3/2} & 0 \\ 0 & 0 & -D^-_{j+1/2} & \ddots & \ddots \\ \ddots & 0 & 0 & \ddots & \ddots \end{pmatrix}$$

To obtain Harten's TVD criteria for the implicit scheme, we need only require that this operator be an M-matrix to obtain the following conditions as did Harten

$$D^+_{j+1/2} \geq 0$$

$$D^-_{j+1/2} \geq 0.$$

2.2.1 Maximum Principles and Monotonicity Preserving Schemes on Multidimensional Structured Meshes

Unfortunately, two motivations suggest a further weakening the concept of monotonicity. The first motivation concerns a negative result by Goodman and Le Veque [GV85] that conservative TVD schemes on Cartesian meshes in two space dimensions are first order accurate. The second motivation is the apparent difficulty in extending the TVD concept to arbitrary unstructured meshes. The first motivation inspired Spekreijse [Spe87] to consider a new class of monotonicity preserving schemes based on positivity of coefficients. Consider the following conservation law equation in two space dimensions

$$u_t + (f(u))_x + (g(u))_y = 0. \tag{2.21}$$

Next construct a discretization of (2.21) on a logically rectangular mesh

$$\begin{aligned} \frac{U^{n+1}_{j,k} - U^n_{j,k}}{\Delta t} &= A^+_{j+\frac{1}{2},k}(U^n_{j+1,k} - U^n_{j,k}) + A^-_{j-\frac{1}{2},k}(U^n_{j-1,k} - U^n_{j,k}) \\ &+ B^+_{j,k+\frac{1}{2}}(U^n_{j,k+1} - U^n_{j,k}) + B^-_{j,k-\frac{1}{2}}(U^n_{j,k-1} - U^n_{j,k}) \end{aligned} \tag{2.22}$$

with *nonlinear* coefficients

$$A^{\pm}_{j+\frac{1}{2},k} = A(\ldots, U^n_{j-1,k}, U^n_{j,k}, U^n_{j+1,k}, \ldots)$$

$$B^{\pm}_{j-\frac{1}{2},k} = B(\ldots, U^n_{j-1,k}, U^n_{j,k}, U^n_{j+1,k}, \ldots)$$

Theorem 2.2.3 The scheme (2.22) exhibits a discrete maximum principle at steady-state if all coefficients are uniformly bounded and nonnegative

$$A^{\pm}_{j\pm\frac{1}{2},k} \geq 0 \quad B^{\pm}_{j\pm\frac{1}{2},k} \geq 0.$$

Furthermore, the scheme (2.22) is monotonicity preserving under a CFL-like condition if

$$\Delta t \leq \min_{\forall j\,k} \frac{1}{\sum_{\pm}(A^{\pm}_{j\pm\frac{1}{2},k} + B^{\pm}_{j,k\pm\frac{1}{2}})}.$$

Proof: The first task is to prove a discrete maximum principle at steady-state by solving for the value at (j,k).

$$\begin{aligned} U_{j,k} &= \frac{\sum_{\pm}(A_{j\pm\frac{1}{2},k}U_{j\pm1,k} + B_{j,k\pm\frac{1}{2}}U_{j,k\pm1})}{\sum_{\pm}(A_{j\pm\frac{1}{2},k} + B_{j,k\pm\frac{1}{2}})} \\ &= \sum_{\pm}(\alpha_{j\pm\frac{1}{2},k}U_{j\pm1,k} + \beta_{j,k\pm\frac{1}{2}}U_{j,k\pm1}) \end{aligned} \tag{2.23}$$

with the constraints

$$\alpha_{j-\frac{1}{2},k} + \alpha_{j+\frac{1}{2},k} + \beta_{j,k-\frac{1}{2}} + \beta_{j,k+\frac{1}{2},k} = 1$$

and $\alpha_{j\pm\frac{1}{2},k} \geq 0$, and $\beta_{j,k\pm\frac{1}{2}} \geq 0$. From positivity of coefficients and convexity of (2.23) it follows that

$$\min(U_{j\pm 1,k}, U_{j,k\pm 1}) \leq U_{j,k} \leq \max(U_{j\pm 1,k}, U_{j,k\pm 1}). \tag{2.24}$$

If $U_{j,k}$ attains a maximum value M at (j,k) then

$$M = U_{j-1,k} = U_{j+1,k} = U_{j,k-1} = U_{j,k+1}.$$

Repeated application of (2.24) to neighboring mesh points establishes the maximum principle.

Next it is straightforward to establish a CFL-like condition for monotonicity preservation in time by again seeking positivity of coefficients and a convex local mapping from U^n to U^{n+1}.

$$\begin{aligned} U^{n+1}_{j,k} &= \left(1 - \Delta t \sum_{\pm}(A^{\pm}_{j\pm\frac{1}{2},k} + B^{\pm}_{j,k\pm\frac{1}{2}})\right) U^n_{j,k} + \Delta t \sum_{\pm}(A_{j\pm\frac{1}{2},k} U^n_{j\pm 1,k} + B_{j,k\pm\frac{1}{2}} U^n_{j,k\pm 1}) \\ &= \gamma_{j,k} U^n_{j,k} + \sum_{\pm}(\alpha_{j\pm\frac{1}{2},k} U^n_{j\pm 1,k} + \beta_{j,k\pm\frac{1}{2}} U^n_{j,k\pm 1}) \end{aligned} \tag{2.25}$$

with the derivable constraints

$$\gamma_{j,k} + \alpha_{j-\frac{1}{2},k} + \alpha_{j+\frac{1}{2},k} + \beta_{j,k-\frac{1}{2}} + \beta_{j,k+\frac{1}{2},k} = 1$$

and $\alpha_{j\pm\frac{1}{2},k} \geq 0$, and $\beta_{j,k\pm\frac{1}{2}} \geq 0$. To show that (2.25) is a local convex mapping, it suffices to satisfy the CFL-like condition for nonnegativity of $\gamma_{j,k}$:

$$\Delta t \leq \min_{\forall j k} \frac{1}{\sum_{\pm}(A^{\pm}_{j\pm\frac{1}{2},k} + B^{\pm}_{j,k\pm\frac{1}{2}})}. \tag{2.26}$$

If (2.26) is satisfied then monotonicity preservation in time follows immediately:

$$\min(U^n_{j\pm 1,k}, U^n_{j,k\pm 1}, U^n_{j,k}) \leq U^{n+1}_{j,k} \leq \max(U^n_{j\pm 1,k}, U^n_{j,k\pm 1}, U^n_{j,k}).$$

■

2.2.2 Maximum Principles and Monotonicity Preserving Schemes on Unstructured Meshes

This section examines the maximum principle theory for conservation laws on unstructured meshes. Specifically, our primary attention focuses on Godunov-like upwind finite volume schemes [God59] utilizing solution reconstruction and evolution. Some early maximum principle results for upwind finite volume schemes can be found in [DD88] [BJ89] [Bar91]. Note that many of these results were subsequently used in implementations of the discontinuous Galerkin method as well, see for example Bey [Bey91]. Also note that the present analysis differs from maximum principle theory based on the "upwind triangle" scheme developed by Desideri and Dervieux [DD88], Arminjon and Dervieux [AD93].

Consider the integral conservation law form of (2.21) for some domain Ω comprised of nonoverlapping control volumes, Ω_i, such that $\Omega = \cup\Omega_i$ and $\Omega_i \cap \Omega_j = 0, i \neq j$. Next, impose the integral conservation law statement on each control volume

$$\frac{\partial}{\partial t}\int_{\Omega_i} u \; d\Omega + \int_{\partial\Omega_i} (F\cdot\mathbf{n}) \; d\Gamma = 0 \tag{2.27}$$

where $F(u) = f(u)\hat{i} + g(u)\hat{j}$. The situation is depicted for a control volume Ω_0 in Fig. 2.5. For two- and three-dimensional triangulations, several control volume choices are available:

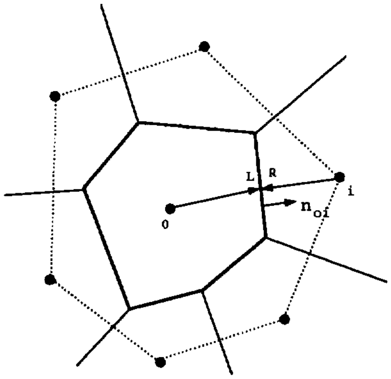

Figure 2.5: Local control volume configuration for unstructured mesh.

the triangles themselves, Voronoi duals, median duals, etc. Although the actual choice of control volume tessellation is very important, the monotonicity analysis contained in the remainder of this section is largely independent of this choice. Consequently, a generic control volume Ω_0 with neighboring control volumes Ω_i, $i \in \mathcal{N}_0$, as shown in Fig. 2.5, is sufficient for the present analysis. In the following example, the solution data is assumed constant in each control volume. This simplifies the analysis considerably. The second example addresses the more general situation utilizing high order data reconstruction.

Example: Analysis of an Upwind Finite Volume Scheme with Piecewise Constant Reconstruction

In this example, assume that the solution data $u(x,y)_i$ in each control volume Ω_i is constant with value equal to the integral average value, i.e.

$$u(x,y)_i = \overline{u}_i = \frac{1}{A_i}\int_{\Omega_i} u \; d\Omega, \;\; \forall\Omega_i \in \Omega$$

where A_i is the area of Ω_i. Next, define the unit exterior normal vector $\mathbf{n}_{0i}$ for the control volume boundary separating Ω_0 and Ω_i. It is also useful to define a normal vector $\vec{\mathbf{n}}_{0i}$ which is scaled by the length (area in 3-D) of that portion of the control volume boundary separating Ω_0 and Ω_i. Finally, to simplify the exposition, define

$$f(u;\vec{\mathbf{n}}) = F(u)\cdot\vec{\mathbf{n}}$$

and assume the existence of a mean value linearization such that

$$f(v;\vec{\mathbf{n}}) - f(u;\vec{\mathbf{n}}) = df(u,v;\vec{\mathbf{n}})(v-u). \tag{2.28}$$

Using this notation, construct the following upwind scheme

$$\frac{d}{dt}(A_0\overline{u}_0) = -\sum_{i\in\mathcal{N}_0} h(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i}) \tag{2.29}$$

with

$$h(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i}) = \frac{1}{2}\left(f(\overline{u}_0;\vec{\mathbf{n}}_{0i}) + f(\overline{u}_{;}\vec{\mathbf{n}}_{0i})\right) - \frac{1}{2}|df(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})|(\overline{u}_i - \overline{u}_0)$$

In Barth and Jespersen [BJ89], we proved a maximum principle and monotonicity preservation of the scheme (2.29) for scalar advection.

Theorem 2.2.4 The upwind algorithm (2.29) with piecewise constant solution data exhibits a discrete maximum principle for arbitrary unstructured meshes and is monotonicity preserving under the CFL-like condition:

$$\Delta t \le \min_{\forall\Omega_j\in\Omega} \frac{-A_j}{\sum_{i\in\mathcal{N}_j} df^-(\overline{u}_i,\overline{u}_j;\vec{\mathbf{n}}_{ji})}.$$

Proof: Consider the control volume surrounding v_0 as shown in Fig. 2.5. Recall that the flux function was constructed using a mean value linearization such that

$$f(\overline{u}_i;\vec{\mathbf{n}}_{0i}) - f(\overline{u}_0;\vec{\mathbf{n}}_{0i}) = df(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})\,(\overline{u}_i - \overline{u}_0)$$

This permits regrouping terms into the following form:

$$\frac{d}{dt}(\overline{u}_0A_0) = -\sum_{i\in\mathcal{N}_0} f(\overline{u}_0;\vec{\mathbf{n}}_{0i}) - \sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})(\overline{u}_i - \overline{u}_0)$$

where $(\cdot) = (\cdot)^+ + (\cdot)^-$ and $|(\cdot)| = (\cdot)^+ - (\cdot)^-$. For any closed control volume, it follows that

$$\sum_{i\in\mathcal{N}_0} f(\overline{u}_0;\vec{\mathbf{n}}_{0i}) = 0.$$

Combining the remaining terms yields a final form for analysis

$$\frac{d}{dt}(\overline{u}_0A_0) = -\sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})(\overline{u}_i - \overline{u}_0). \tag{2.30}$$

To verify a maximum principle at steady-state, set the time term to zero and solve for $\overline{u}_0$.

$$\overline{u}_0 = \frac{\sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})\;\overline{u}_i}{\sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0,\overline{u}_i;\vec{\mathbf{n}}_{0i})} = \sum_{i\in\mathcal{N}_0} \alpha_i\overline{u}_i$$

with $\sum_{i\in\mathcal{N}_0}\alpha_i = 1$ and $\alpha_i \ge 0$. Since $\overline{u}_0$ is a convex combination of all neighbors

$$\min_{i\in\mathcal{N}_0}\overline{u}_i \le \overline{u}_0 \le \max_{i\in\mathcal{N}_0}\overline{u}_i. \tag{2.31}$$

If $\overline{u}_0$ takes on a maximum value M in the interior, then $\overline{u}_i = M, \forall\ i \in \mathcal{N}_0$. Repeated application of (2.31) to neighboring control volumes establishes the maximum principle.

For Euler explicit time stepping,

$$\frac{d}{dt}(\overline{u}_0A_0) \approx A_0\frac{\overline{u}_0^{n+1} - \overline{u}_0^n}{\Delta t},$$

a CFL-like condition is obtained for monotonicity preservation. Inserting this expression into (2.30) yields

$$\begin{aligned}\overline{u}_0^{n+1} &= \overline{u}_0^n - \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0^n,\overline{u}_i^n;\vec{\mathbf{n}}_{0i})(\overline{u}_i^n-\overline{u}_0^n) \\ &= \alpha_0\overline{u}_0^n+\sum_{i\in\mathcal{N}_0}\alpha_i\overline{u}_i^n. \end{aligned} \tag{2.32}$$

It should be clear that coefficients in (2.32) sum to unity. To prove monotonicity preservation, it is sufficient to show nonnegativity of coefficients. Clearly, $\alpha_i \geq 0 \;\; \forall\, i > 0$, hence, monotonicity preservation is achieved if

$$\alpha_0 = 1 + \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0} df^-(\overline{u}_0^n,\overline{u}_i^n;\vec{\mathbf{n}}_{0i}) \;\geq 0.$$

This implies monotonicity preservation in time under the CFL-like condition

$$\Delta t \leq \min_{\forall\Omega_j\in\Omega}\frac{-A_j}{\sum_{i\in\mathcal{N}_j} df^-(\overline{u}_0^n,\overline{u}_i^n;\vec{\mathbf{n}}_{0i})}.$$

■

Example: Analysis of High Order Accurate Upwind Advection Schemes Using Arbitrary Order Reconstruction [Bar94]

In this example, high order accurate upwind schemes on unstructured meshes are considered. The technique used here is to show a maximum principle for the cell averages. The solution algorithm is a relatively standard procedure for extensions of Godunov's scheme in Eulerian coordinates, see for example [God59, vL79, CW84, HOEC87]. The basic idea in Godunov's method is to treat the integral control volume averages, $\overline{u}$, as the basic unknowns. Using information from the control volume averages, $k-th$ order piecewise polynomials are *reconstructed* in each control volume Ω_i:

$$U^k(x,y)_i = \sum_{m+n\leq k}\alpha_{(m,n)}P_{(m,n)}(x-x_c,y-y_c)$$

where $P_{(m,n)}(x-x_c,y-y_c) = (x-x_c)^m(y-y_c)^n$ and (x_c,y_c) is the control volume centroid. The process of reconstruction amounts to finding the polynomial coefficients, $\alpha_{(m,n)}$. Near steep gradients and discontinuities, these polynomial coefficients maybe altered based on monotonicity arguments. Because the reconstructed polynomials vary discontinuously from control volume to control volume, a unique value of the solution does not exist at control volume interfaces. This non-uniqueness is resolved via exact or approximate solutions of the Riemann problem. In practice, this is accomplished by supplanting the true flux function with a numerical flux function which produces a single unique flux given two solution states. Once the flux integral is carried out (either exactly or by numerical quadrature), the control volume average of the solution can be evolved in time. In most cases, standard techniques for integrating ODE equations are used for the time evolution, i.e. Euler implicit, Euler explicit, Runge-Kutta. The result of the evolution process is a new collection of control volume averages. The process can then be repeated. The process can be summarized in the following steps:

(1) **Reconstruction in Each Control Volume**: Given integral solution averages in all Ω_j, reconstruct a $k-th$ order piecewise polynomial $U^k(x,y)_i$ in each Ω_i for use in equation (2.27). In faithful implementations of Godunov's method, cell averages of the solution data

$$\int_{\Omega_i} U^k(x,y)_i \, d\Omega = (\overline{u}A)_i$$

are preserved during the reconstruction process. For solutions containing discontinuities and/or steep gradients, monotonicity enforcement may be required.

(2) **Flux Evaluation on Each Edge**: Supplant the true flux by a numerical flux function. Given two solution states the numerical flux function returns a single unique flux. Using the notation of the previous section, define $f(u;\mathbf{n}) = (F(u)\cdot\mathbf{n})$ so that

$$\int_{\partial\Omega_i} f(u;\mathbf{n}) \, d\Gamma \approx \int_{\partial\Omega_i} h(U^L,U^R;\mathbf{n}) d\Gamma.$$

Consider each control volume boundary $\partial\Omega_i$, to be a collection of polygonal edges (or dual edges) from the mesh. Along each edge (or dual edge), perform a high order accurate flux quadrature. When the reconstruction polynomial is piecewise linear, single (midpoint) quadrature is usually employed on both structured and unstructured meshes

$$\int_{\partial\Omega_i} h(U^L,U^R;\mathbf{n}) d\Gamma \approx \sum_{j\in\mathcal{N}_i} h(U^L,U^R;\vec{\mathbf{n}})_{ij}$$

where U^L and U^R are solution values evaluated at the midpoint of control volume edges as shown in Fig. 2.5. When multi-point quadrature formulas are employed, they are assumed to be of the form:

$$\int_0^1 f(s)ds = \sum_{q\in Q} w_q f(\xi_q)$$

with $w_q > 0$ and $\xi_q \in [0,1]$. Let the multi-point quadrature formulas be represented by the augmented notation

$$\int_{\partial\Omega_i} h(U^L,U^R;\mathbf{n}) d\Gamma \approx \sum_{j\in\mathcal{N}_i} \sum_{q\in Q} w_q h(U^L,U^R;\vec{\mathbf{n}})_{ijq}$$

(3) **Evolution in Each Control Volume**: Collect flux contributions in each control volume and evolve in time using any time stepping scheme, i.e. Euler explicit, Euler implicit, Runge-Kutta, etc. The result of this step is once again control volume averages and the process can be repeated.

In the present analysis, the reconstruction polynomials $U^k(x,y)_i$ in each Ω_i are given. The result of the analysis will be conditions or constraints on the reconstruction so that a maximum principle involving cell averages can be obtained. The topic of reconstruction and implementation of the constraints determined by this analysis will be examined in a later section. Using this notation, the following upwind scheme is constructed for the configuration in Fig. 2.5.

$$\frac{d}{dt}(A_0\overline{u}_0) = -\sum_{i\in\mathcal{N}_0} \sum_{q\in Q} w_q h(U^L,U^R;\vec{\mathbf{n}})_{0iq} \tag{2.33}$$

with a numerical flux function obtained from (2.28).

$$\begin{aligned} h(U^L, U^R; \vec{\mathbf{n}}_{0i}) &= \frac{1}{2}\left(f(U^L;\vec{\mathbf{n}}) + f(U^R;\vec{\mathbf{n}})\right)_{0i} \\ &- \frac{1}{2}|df(U^L,U^R;\vec{\mathbf{n}})|_{0i}(U^R - U^L)_{0i} \end{aligned} \qquad (2.34)$$

To analyze the scheme, recall that the flux function was constructed using a mean value linearization such that

$$f(U^R;\vec{\mathbf{n}}) - f(U^L;\vec{\mathbf{n}}) = df(U^L,U^R;\vec{\mathbf{n}})(U^R - U^L).$$

This permits regrouping terms into the following form:

$$\frac{d}{dt}(\overline{u}_0 A_0) = -\sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q\left(f(U^L;\vec{\mathbf{n}}) + df^-(U^L,U^R;\vec{\mathbf{n}})(U^R - U^L)\right)_{0iq}. \qquad (2.35)$$

Rewrite the first term in the sum using a mean value construction

$$\sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q f(\overline{u}_0;\vec{\mathbf{n}})_{0iq} + w_q\left(df(\overline{u}_0,U^L;\vec{\mathbf{n}})(U^L - \overline{u}_0)\right)_{0iq}.$$

The first term vanishes when summed over a closed volume so that (2.35) reduces to

$$\frac{d}{dt}(\overline{u}_0 A_0) = -\sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q\left(df(\overline{u}_0,U^L;\vec{\mathbf{n}})(U^L - \overline{u}_0) + df^-(U^L,U^R;\vec{\mathbf{n}})(U^R - U^L)\right)_{0iq}$$

By introducing difference ratios, the scheme can be written in the following form:

$$\begin{aligned} \frac{d}{dt}(\overline{u}_0 A_0) &= -\sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q\left(df^-(\overline{u}_0,U^L;\vec{\mathbf{n}})\Psi\right)_{0iq}(\overline{u}_i - \overline{u}_0) \\ &- \sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q\left(df^+(\overline{u}_0,U^L;\vec{\mathbf{n}})\Phi\right)_{0iq}(\overline{u}_0 - \overline{u}_k) \\ &- \sum_{i\in\mathcal{N}_0}\sum_{q\in Q} w_q\left(df^-(U^L,U^R;\vec{\mathbf{n}})\Theta\right)_{0iq}(\overline{u}_i - \overline{u}_0) \end{aligned} \qquad (2.36)$$

with

$$\Psi_{0iq} = \frac{U^L_{0iq} - \overline{u}_0}{\overline{u}_i - \overline{u}_0}, \quad \Phi_{0iq} = \frac{U^L_{0iq} - \overline{u}_0}{\overline{u}_0 - \overline{u}_k}, \quad \Theta_{0iq} = \frac{U^R_{0iq} - U^L_{0iq}}{\overline{u}_i - \overline{u}_0}.$$

In this equation, the k subscript refers to some as yet unspecified index value, $k \in \mathcal{N}_0$.

Theorem 2.2.5 The generalized Godunov scheme with arbitrary order reconstruction (2.33) exhibits a discrete maximum principle at steady-state if the following three conditions are fulfilled:

$$\Psi_{jiq} \geq 0, \quad \Phi_{jiq} \geq 0 \quad \Theta_{jiq} \geq 0 \;\; \forall j, q, \; i \in \mathcal{N}_j \qquad (2.37)$$

as defined by (2.36). Furthermore, the scheme is monotonicity preserving under a CFL-like condition if

$$\Delta t \leq \min_{\forall\, \Omega_j \in \Omega} \frac{-A_j}{\sum_{i\in\mathcal{N}_j}\sum_{q\in Q}\left(\overline{df}^-\Psi - \overline{df}^+\Phi + df^-\Theta\right)_{jiq}}$$

Proof: Assume that (2.37) holds. Define $\overline{df}_{0iq} = w_q df(\overline{u}_0, U^L; \vec{\mathbf{n}})_{0iq}$ and similarly $df_{0iq} = w_q df(U^L, U^R; \vec{\mathbf{n}})_{0iq}$. Setting the time term to zero and solving for $\overline{u}_0$ yields

$$\begin{aligned} \overline{u}_0 &= \frac{\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\left(\overline{df}^{+}\Phi\right)_{0iq}\overline{u}_k - \left(\overline{df}^{-}\Psi + df^{-}\Theta\right)_{0iq}\overline{u}_i}{\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\left(\overline{df}^{+}\Phi - \overline{df}^{-}\Psi - df^{-}\Theta\right)_{0iq}} \\ &= \sum_{i\in\mathcal{N}_0} \alpha_i \overline{u}_i. \end{aligned} \tag{2.38}$$

Examining the individual coefficients, it is clear that $\sum_{i\in\mathcal{N}_0} \alpha_i = 1$ and $\alpha_i \geq 0, \forall i$. Thus a convex local mapping exists and

$$\min_{i\in\mathcal{N}_0} \overline{u}_i \leq \overline{u}_0 \leq \max_{i\in\mathcal{N}_0} \overline{u}_i. \tag{2.39}$$

If $\overline{u}_0$ takes on a maximum value M in the interior, then $\overline{u}_i = M, \forall\ i \in \mathcal{N}_0$. Repeated application of (2.39) to neighboring control volumes establishes the maximum principle.

To establish monotonicity preservation in time, consider Euler explicit time-stepping scheme.

$$\frac{d}{dt}(\overline{u}_0 A_0) \approx A_0 \frac{\overline{u}_0^{n+1} - \overline{u}_0^n}{\Delta t}$$

Inserting this formula into (2.36) yields

$$\begin{aligned} \overline{u}_0^{n+1} &= \overline{u}_0^n - \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\overline{df}^{-}_{0iq}\Psi_{0iq}(\overline{u}_i^n - \overline{u}_0^n) \\ &- \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\overline{df}^{+}_{0iq}\Phi_{0iq}(\overline{u}_0^n - \overline{u}_k^n) \\ &- \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}df^{-}_{0iq}\Theta_{0iq}(\overline{u}_i^n - \overline{u}_0^n) \\ &= \alpha_0\overline{u}_0^n + \sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\alpha_i\overline{u}_i^n \end{aligned} \tag{2.40}$$

with $\alpha_0 + \sum_{i\in\mathcal{N}_0} \alpha_i = 1$ and $\alpha_i \geq 0, i > 0$. A locally convex mapping in time from U^n to U^{n+1} is achieved when $\alpha_0 \geq 0$. This assures monotonicity in time. Some algebra reveals the following formula for α_0

$$\alpha_0 = 1 + \frac{\Delta t}{A_0}\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\left(\overline{df}^{-}\Psi - \overline{df}^{+}\Phi + df^{-}\Theta\right)_{0iq}$$

From this the CFL-like condition for monotonicity preservation is obtained

$$\Delta t \leq \min_{\forall\ \Omega_i\in\Omega} \frac{-A_0}{\sum_{i\in\mathcal{N}_0}\sum_{q\in Q}\left(\overline{df}^{-}\Psi - \overline{df}^{+}\Phi + df^{-}\Theta\right)_{0iq}}$$

so that

$$\min_{i\in\mathcal{N}_0}(\overline{u}_i^n, \overline{u}_0^n) \leq \overline{u}_0^{n+1} \leq \max_{i\in\mathcal{N}_0}(\overline{u}_i^n, \overline{u}_0^n)$$

Applying this result to all control volumes establishes monotonicity preservation. ■

Without specifying the actual type of reconstruction, we have the following simple lemma concerning the ratios appearing in (2.36):

Lemma 2.2.1 Assume $\Psi_{0iq} \geq 0$ as defined in (2.36). A sufficient condition for $\Phi_{0iq} \geq 0$ is that the reconstructed polynomial reduce to the cell average value, $U^k(x,y)_0 = \overline{u}_0$, at local cell average extrema, i.e. whenever

$$\max_{j \in \mathcal{N}_0} \overline{u}_j \leq \overline{u}_0 \leq \min_{j \in \mathcal{N}_0} \overline{u}_j.$$

Proof: Consider an arbitrary control volume Ω_i adjacent to Ω_0. Assume that $\overline{u}_i \geq \overline{u}_0$. The stated assumption, $\Psi_{0iq} \geq 0$ implies that $U^L_{0iq} \geq \overline{u}_0$. Consequently, $\Phi_{0iq} \leq 0$ if and only if $\overline{u}_0$ is less than all adjacent neighbors, hence a local minimum. Following a similar argument, $\overline{u}_0$ is a local maximum when $\overline{u}_i \leq \overline{u}_0$ and $\Phi_{0iq} \leq 0$. ■

Chapter 3

Upwind Finite Volume Schemes

This chapter examines upwind finite volumes schemes for scalar and system conservation laws. The basic tasks in the upwind finite volume approach have already been presented: reconstruction, flux evaluation, and evolution. By far, the most difficult task in this process is the reconstruction step.

3.1 Reconstruction Schemes for Upwind Finite Volume Schemes

In the following paragraphs, the design criteria for general reconstruction operators with fixed stencil is reviewed. The reader is referred to the papers by Abgrall [Abg94, Abg95], Vankeirsbilck [Van93] and Michell [Mic94] for a discussion of ENO and adaptive stencil reconstruction schemes.

The reconstruction operator serves as a finite-dimensional (possibly pseudo) inverse of the cell-averaging operator $\mathbf{A}$ whose j-th component $\mathbf{A}_j$ computes the cell average of the solution in Ω_j.

$$\overline{u}_j = \mathbf{A}_j u = \frac{1}{A_j} \int_{\Omega_j} u(x,y)\, d\Omega$$

In addition, the following properties are usually imposed on the reconstruction:

(1) **Conservation of the mean**: Simply stated, given cell averages $\overline{u}$, we require that all polynomial reconstructions u^k have the correct cell average.

$$\text{if } \; u^k = \mathbf{R}^k \overline{u} \; \text{ then } \; \overline{u} = \mathbf{A} u^k$$

This means that $\mathbf{R}^k$ is a right inverse of the averaging operator $\mathbf{A}$.

$$\mathbf{A}\mathbf{R}^k = I$$

Conservation of the mean has an important implication. Unlike finite-element schemes, *Godunov schemes have a diagonal mass matrix.*

(2) **k-exactness**: A reconstruction operator $\mathbf{R}^k$ is *k-exact* if $\mathbf{R}^k\mathbf{A}$ reconstructs polynomials of degree k or less exactly.

$$\text{if } \; u \in \mathcal{P}_k \; \text{ and } \; \overline{u} = \mathbf{A}u, \; \text{ then } \; u^k = \mathbf{R}^k \overline{u} = u$$

In other words, $\mathbf{R}^k$ is a left-inverse of $\mathbf{A}$ restricted to the space of polynomials of degree at most k.

$$\mathbf{R}^k \mathbf{A}\Big|_{\mathcal{P}_k} = I$$

This insures that exact solutions contained in $\mathcal{P}_k$ are in fact solutions of the discrete equations. For sufficiently smooth solutions, the property of k-exactness also insures that when piecewise polynomials are evaluated at control volume boundaries, the difference between solution states diminishes with increasing k at a rate proportional to h^{k+1} were h is a maximum diameter of the two control volumes. Figure 3.1 shows a global quartic polynomial $u \in \mathcal{P}_4$ which has been averaged in each interval. This same figure shows linear and quadratic reconstruction given interval averages. The small jumps in the piecewise polynomials at interval boundaries would decrease even more for cubics and vanish altogether for quartic reconstruction. Property (1) requires that the area under each piecewise polynomial is exactly equal to the cell average.

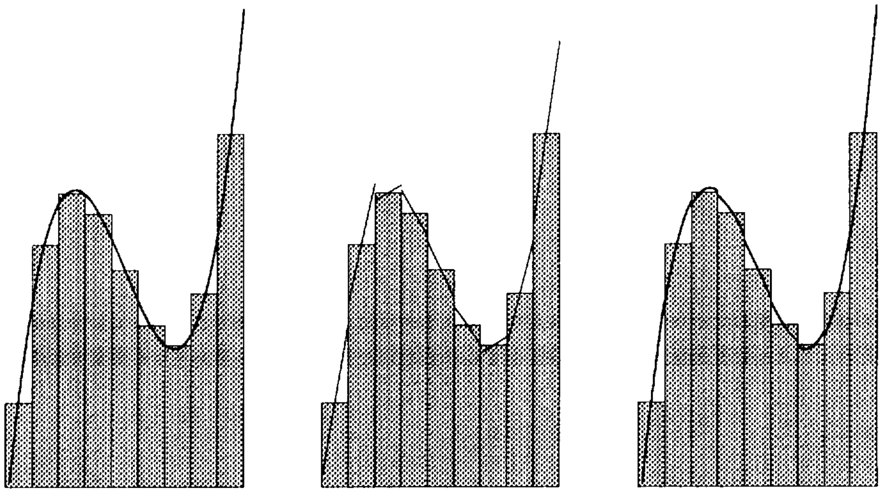

Figure 3.1: Cell averaging of quartic polynomial (left), linear reconstruction (center) and quadratic reconstruction (right).

One of the most important observations concerning linear reconstruction is that one can dispense with the notion of cell averages as unknowns by reinterpreting the unknowns as pointwise values of the solution sampled at the centroid (midpoint in 1-D) of the control volume. This well known result greatly simplifies schemes based on linear reconstruction. The linear reconstruction in each interval shown in Fig. 3.1 was obtained by a simple central-difference formula given pointwise values of the solution at the midpoint of each interval. Note that for steady-state computations, conservation of the mean in the data reconstruction is not necessary. The implication of violating this conservation is that a *nondiagonal* mass matrix appears in the time integral. Since time derivatives vanish at steady-state, the effect of this mass matrix vanishes at steady-state. The reconstruction schemes presented below assume that solution variables are placed at the vertices of the mesh, which may not be at the precise centroid.

3.1.1 Green-Gauss Reconstruction

Let D_0 denote the set of all triangles incident to some vertex v_0 and the exact integral relation

$$\int_{D_0} \nabla u \, d\Omega = \int_{\partial D_0} u \, \mathbf{n} \, d\Gamma. \tag{3.1}$$

It is not difficult to show [BJ89] that given function values at vertices of a triangulation, a discretization of this formula can be constructed which is exact whenever u varies linearly:

$$(\nabla u)_{v_0} = \frac{1}{A_0} \sum_{i \in \mathcal{N}_0} \frac{1}{2}(u_i + u_0)\vec{\mathbf{n}}_{0i}. \tag{3.2}$$

In this formula $\vec{\mathbf{n}}_{0i} = \int_a^b d\,\vec{\mathbf{n}}$ for any path which connects triangle centroids adjacent to the edge $e(v_0, v_i)$ and A_0 is the area of the *nonoverlapping* dual regions formed by this choice of path integration. Two typical choices are the median and centroid duals as shown below. This approximation extends naturally to three dimensions. The formula (3.2) suggests a natural computer implementation using the edge data structure. Assume that the normals $\vec{\mathbf{n}}_{ij}$ for all edges $e(v_i, v_j)$ have been precomputed with the convention that the normal vector points from v_i to v_j. An edge implementation of (3.2) can be performed in the following way:

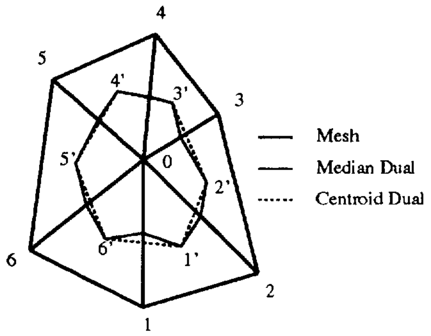

Figure 3.2: Local mesh with centroid and median duals.

```
For $k = 1, n(e)$   ! Loop through edges of mesh
    $j_1 = e^{-1}(k, 1)$  !Pointer to edge origin
    $j_2 = e^{-1}(k, 2)$  !Pointer to edge destination
    $uav = (u(j_1) + u(j_2))/2$  !Gather
    $ux(j_1) \mathrel{+}= normx(k) \cdot uav$  !Scatter
    $ux(j_2) \mathrel{-}= normx(k) \cdot uav$
    $uy(j_1) \mathrel{+}= normy(k) \cdot uav$
    $uy(j_2) \mathrel{-}= normy(k) \cdot uav$
Endfor
For $j = 1, n(v)$   ! Loop through vertices
    $ux(j) = ux(j)/area(j)$   ! Scale by area
    $uy(j) = uy(j)/area(j)$
Endfor
```

It can be shown that the use of edge formulas for the computation of vertex gradients is asymptotically optimal in terms of work done.

3.1.2 Linear Least-Squares Reconstruction

To derive this reconstruction technique, consider a vertex v_0 and suppose that the solution varies linearly over the support of adjacent neighbors of the mesh. In this case, the change in vertex values of the solution along an edge $e(v_i, v_0)$ can be calculated by

$$(\nabla u)_0 \cdot (\mathbf{r}_i - \mathbf{r}_0) = u_i - u_0.$$

This equation represents the scaled projection of the gradient along the edge $e(v_i, v_0)$. A similar equation could be written for all incident edges subject to an arbitrary weighting factor. The result is the following matrix equation, shown here in two dimensions:

$$\begin{bmatrix} w_1\Delta x_1 & w_1\Delta y_1 \\ \vdots & \vdots \\ w_n\Delta x_n & w_n\Delta y_n \end{bmatrix} \begin{pmatrix} u_x \\ u_y \end{pmatrix} = \begin{pmatrix} w_1(u_1-u_0) \\ \vdots \\ w_n(u_n-u_0) \end{pmatrix}$$

or in symbolic form $\mathcal{L}\ \nabla u = \mathbf{f}$ where

$$\mathcal{L} = [\, \vec{L}_1 \quad \vec{L}_2 \,]$$

in two dimensions. Exact calculation of gradients for linearly varying u is guaranteed if any two row vectors $w_i(\mathbf{r}_i - \mathbf{r}_0)$ span all of 2 space. This implies linear independence of $\vec{L}_1$ and $\vec{L}_2$. The system can then be solved via a Gram-Schmidt process, i.e.,

$$\begin{bmatrix} \vec{V}_1 \\ \vec{V}_2 \end{bmatrix} [\, \vec{L}_1 \quad \vec{L}_2 \,] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{5.3.4}$$

The row vectors $\vec{V}_i$ are given by

$$\begin{aligned} \vec{V}_1 &= \frac{l_{22}\vec{L}_1 - l_{12}\vec{L}_2}{l_{11}l_{22} - l_{12}^2} \\ \vec{V}_2 &= \frac{l_{11}\vec{L}_2 - l_{12}\vec{L}_1}{l_{11}l_{22} - l_{12}^2} \end{aligned} \tag{3.3}$$

with $l_{ij} = (\vec{L}_i \cdot \vec{L}_j)$.

Note that reconstruction of N independent variables in $\mathbf{R}^d$ implies $\binom{d+1}{2} + d\,N$ inner product sums. Since only $d\,N$ of these sums involves the solution variables themselves, the remaining sums could be precalculated and stored in computer memory. This makes the present scheme competitive with the Green-Gauss reconstruction. Using the edge data structure, the calculation of inner product sums can be calculated for *arbitrary* combinations of polyhedral cells. In all cases linear functions are reconstructed exactly. This technique is best illustrated by example:

```
For $k = 1, n(e)$  !Loop through edges of mesh
    $j_1 = e^{-1}(k,1)$  !Pointer to edge origin
    $j_2 = e^{-1}(k,2)$  !Pointer to edge destination
    $dx = w(k) \cdot (x(j_2) - x(j_1))$  !Weighted $\Delta x$
    $dy = w(k) \cdot (y(j_2) - y(j_1))$  !Weighted $\Delta y$
    $l_{11}(j_1) = l_{11}(j_1) + dx \cdot dx$  ! $l_{11}$ orig sum
    $l_{11}(j_2) = l_{11}(j_2) + dx \cdot dx$  ! $l_{11}$ dest sum
    $l_{12}(j_1) = l_{12}(j_1) + dx \cdot dy$  ! $l_{12}$ orig sum
    $l_{12}(j_2) = l_{12}(j_2) + dx \cdot dy$  ! $l_{12}$ dest sum
    $du = w(k) \cdot (u(j_2) - u(j_1))$  !Weighted $\Delta u$
    $lf_1(j_1) \mathrel{+}= dx \cdot du$  !$\vec{L}_1 f$ sum
    $lf_1(j_2) \mathrel{+}= dx \cdot du$
    $lf_2(j_1) \mathrel{+}= dy \cdot du$  !$\vec{L}_2 f$ sum
    $lf_2(j_2) \mathrel{+}= dy \cdot du$
Endfor
For $j = 1, n(v)$  ! Loop through vertices
```

```
    $det = l_{11}(j) \cdot l_{22}(j) - l_{12}^2$
    $ux(j) = (l_{22}(j) \cdot lf_1(j) - l_{12} \cdot lf_2)/det$
    $uy(j) = (l_{11}(j) \cdot lf_2(j) - l_{12} \cdot lf_1)/det$
Endfor
```

This formulation provides freedom in the choice of weighting coefficients, w_i. These weighting coefficients can be a function of the geometry and/or solution. Classical approximations in one dimension can be recovered by choosing geometrical weights of the form $w_i = 1./|\Delta\mathbf{r}_i - \Delta\mathbf{r}_0|^t$ for values of $t = 0, 1, 2$.

3.1.3 Monotonicity Enforcement

When solution discontinuities and steep gradients as present, additional steps must be taken to prevent oscillations from developing in the numerical solution. One way to do this was pioneered by van Leer [vL79] in the late 1970's. His basic idea was to enforce strict monotonicity in the reconstruction. Monotonicity in this context means that the value of the reconstructed polynomial does not exceed the minimum and maximum of neighboring cell averages. The final reconstruction must guarantee that no new extrema have been created. When a new extremum is produced, the slope of the reconstruction in that interval is reduced until monotonicity is restored. This implies that at a local minimum or maximum in the cell averaged data the slope in 1-D is *always* reduced to zero, see for example Fig. 3.3. Theorem 2.2.5 provides sufficient conditions for a discrete maximum principle in the cell averages using arbitrary order reconstruction on general control volumes. Consider the control volume interface separating Ω_0 and Ω_i as shown in Fig. 2.5. From Theorem 2.2.5, a maximum principle is guaranteed if for all quadrature points on the interface separating Ω_0 and Ω_i

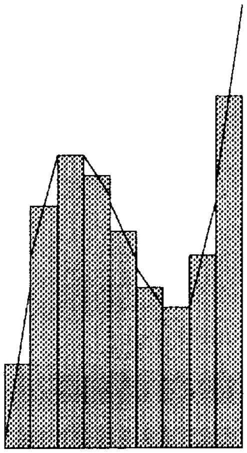

Figure 3.3: Linear data reconstruction with monotone limiting.

$$\Psi_{0i} \geq 0, \quad \Phi_{0i} \geq 0 \quad \Theta_{0i} \geq 0.$$

Lemma 2.2.1 states that $\Phi_{0i} \geq 0$ is always satisfied if the monotonicity enforcement algorithm reduces to piecewise constant at local extremum, i.e. when

$$\max_{j \in \mathcal{N}_0} \overline{u}_j \leq \overline{u}_0 \leq \min_{j \in \mathcal{N}_0} \overline{u}_j.$$

Assume that this property holds, monotonicity reduces to the following two conditions at all quadrature points:

$$\begin{array}{ll} 0 \leq \frac{U^L - \overline{u}_0}{\overline{u}_i - \overline{u}_0} & (a) \\ 0 \leq \frac{U^R - U^L}{\overline{u}_i - \overline{u}_0} & (b) \end{array} \tag{3.4}$$

The second inequality appearing in (3.4) requires that the difference in the extrapolated states at a cell interface must be of the same sign as the difference in the cell average values. For example in Fig. 3.4(a) this condition is violated but can be remedied either by a symmetric reduction of slopes or by replacing the larger slope by the minimum value of the two slopes. *Observe that in one space dimension the net effect of the slope limiting in the reconstruction process is to ensure that the total variation of the reconstructed function does not exceed the total variation of the cell averaged data.*

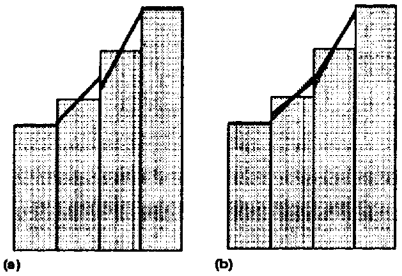

Figure 3.4: (a) Reconstruction profile with increased variation violating monotonicity constraints. (b) Profile after modification to satisfy monotonicity constraints.

In Barth and Jespersen [BJ89], a simple recipe was proposed for slope limiting of linearly reconstructed data on arbitrary unstructured meshes. Consider writing the linearly reconstructed data in the following form for Ω_0:

$$U(x,y)_0 = \overline{u}_0 + \nabla u_0 \cdot (\mathbf{r} - \mathbf{r}_0).$$

Now consider a "limited" form of this piecewise linear distribution.

$$U(x,y)_0 = \overline{u}_0 + \Phi_0 \nabla u_0 \cdot (\mathbf{r} - \mathbf{r}_0)$$

$$\begin{aligned} u_0^{min} &= \min_{i \in \mathcal{N}_0}(\overline{u}_0, \overline{u}_i) \\ u_0^{max} &= \max_{i \in \mathcal{N}_0}(\overline{u}_0, \overline{u}_i) \end{aligned}$$

and require that

$$u_0^{min} \leq U(x,y)_0 \leq u_0^{max}$$

when evaluated at the quadrature points used in the flux integral computation. For each quadrature point location in the flux integral, compute the extrapolated state U_{0i}^L and determine the smallest Φ_0 so that

$$\Phi_0 = \begin{cases} \min(1, \frac{u_0^{max} - \overline{u}_0}{U_{0i}^L - \overline{u}_0}), & \text{if } U_{0i}^L - \overline{u}_0 > 0 \\ \min(1, \frac{u_0^{min} - \overline{u}_0}{U_{0i}^L - \overline{u}_0}), & \text{if } U_{0i}^L - \overline{u}_0 < 0 \\ 1 & \text{if } U_{0i}^L - \overline{u}_0 = 0 \end{cases} \cdot$$

This limiter function automatically satisfies Lemma 2.2.1. Condition (a) from (3.4) is local to the control volume and can be enforced by a further reduction in slope. In practice this step is sometimes omitted. Condition (b) is enforced using the procedure shown in Fig. 3.4.

Extensive numerical testing has shown that this limiter can noticeably degrade the overall accuracy of computations, especially for flows on coarse meshes. In addition the limiter behaves poorly when the solution is nearly constant unless additional heuristic parameters are added. This has prompted other researchers (c.f. Venkatakrishnan [Ven93]) to proposed alternative "smooth" limiter functions, but no serious attempt is made to appeal to the rigors of maximum principle theory. The design of accurate limiters satisfying the maximum principle constraints is still an open problem in this area.

When the above procedures are combined with the flux function given earlier (2.34),

$$\begin{aligned} h(U^L,U^R;\mathbf{n}) &= \frac{1}{2}\left(f(U^L;\mathbf{n}) + f(U^R;\mathbf{n})\right) \\ &- \frac{1}{2}|df(U^L,U^R;\mathbf{n})|\left(U^R - U^L\right) \end{aligned} \tag{3.5}$$

the resulting scheme has good shock resolving characteristics. To demonstrate this, we consider the scalar nonlinear hyperbolic problem suggested by Struijs, Deconinck, *et al.* [Str89]. The equation is a multidimensional form of Burger's equation.

$$u_t + (u^2/2)_x + u_y = 0$$

This equation is solved in a square region $[0, 1.5] \times [0, 1.5]$ with boundary conditions: $u(x,0) = 1.5 - 2x$, $x \leq 1$, $u(x,0) = -.5$, $x > 1$, $u(0,y) = 1.5$, and $u(1.5,y) = -.5$. Figure 3.5 shows carpet plots and contours of the solution on regular and irregular meshes. The exact solution to this problem consists of converging straightline characteristics which eventually form a shock which propagates to the upper boundary. The carpet plots indicate that the numerical solution on both meshes is monotone. Even so, most people would prefer the solution on the regular mesh. This is an unavoidable consequence of irregular meshes. The only remedy appears to be mesh adaptation.

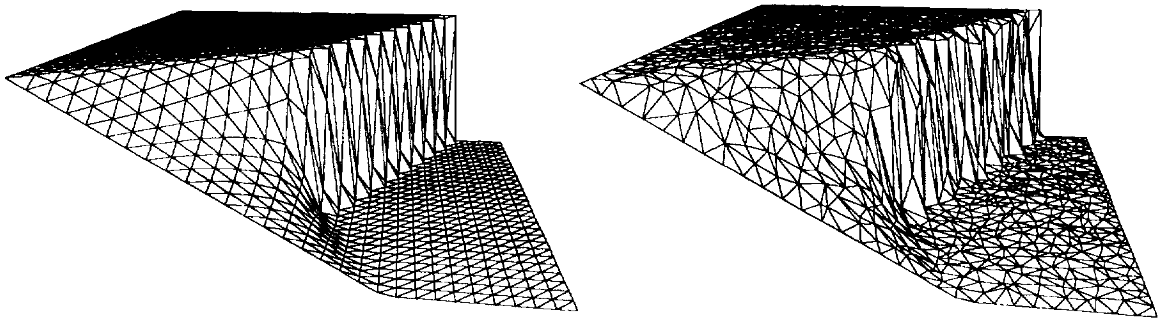

Figure 3.5: Carpet plot of Burger's equation solution on (a) regular mesh, (b) irregular mesh.

Next, consider a test problem which solves the two-dimensional scalar advection equation

$$u_t + (yu)_x - (xu)_y = 0$$

or equivalently

$$u_t + \vec{\lambda} \cdot \nabla u = 0, \quad \vec{\lambda} = (y, -x)^T$$

on a grid centered about the origin, see Fig. 3.6. Discontinuous inflow data is specified along an interior cut line, $u(x,0) = 1$ for $-.6 < x < -.3$ and $u(x,0) = 0$, otherwise. The exact solution is a solid body rotation of the cut line data throughout the domain. The discontinuities admitted by this equation are similar to the linear contact and slip line solutions admitted by the Euler equations. Figure 3.6 also shows solution contours obtained using piecewise constant reconstruction. The discontinuities are quickly smeared by the computation. Figure 3.7 displays solution contours obtained using piecewise linear and a piecewise quadratic reconstruction technique discussed in [Bar93, Bar94]. The improvement from piecewise constant reconstruction to piecewise linear is quite dramatic. The improvement from piecewise linear to piecewise quadratic also looks impressive. The width of the

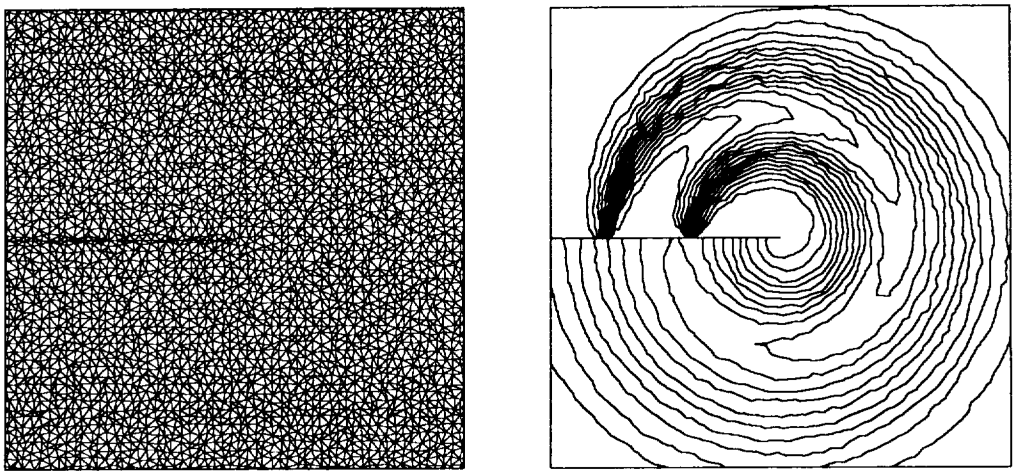

Figure 3.6: Grid for the circular advection problem (left) and solution contours obtained using piecewise constant reconstruction (right).

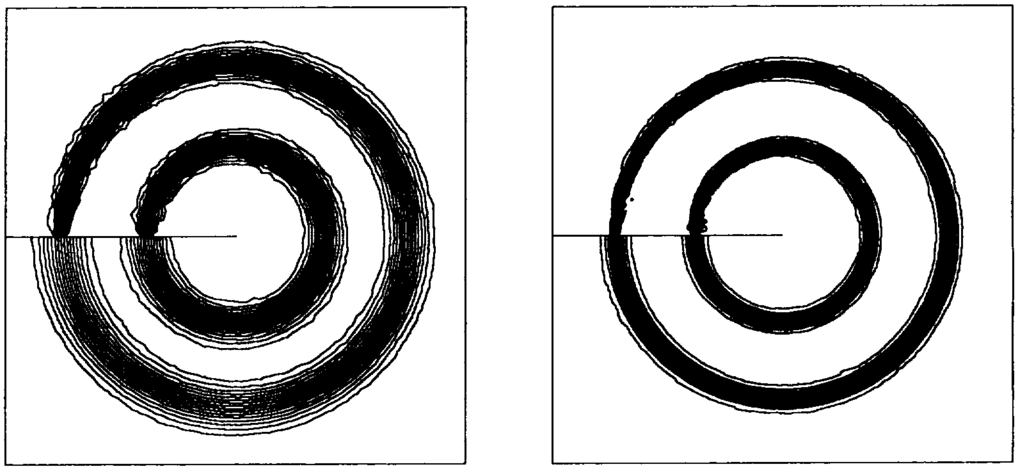

Figure 3.7: Solution contours, piecewise linear reconstruction (left) and piecewise quadratic reconstruction (right).

discontinuities is substantially reduced with little observable grid dependence. Note however, that the quadratic approximation used for this computation has roughly quadrupled the number of solution unknowns because of the use of 6-noded triangles.

3.2 Numerical Solution of the Euler and Navier-Stokes Equations Using Upwind Finite Volume Approximation

The section discusses the extension and application of the scalar finite volume scheme to the Euler and Navier-Stokes equations. One advantage of the finite volume method is the relative ease in which this can be done.

3.2.1 Extension of Scalar Advection Schemes to Systems of Equations

The extension of the scalar advection schemes to the Euler (and Navier-Stokes) equations requires three rather minor modifications:

1. *Vector Flux Function.* The scalar flux function is replaced by a vector flux function. In the present work, the mean value linearization due to Roe [Roe81] is used. The form of this vector flux function is identical to the scalar flux function (2.34), i.e.

$$\begin{aligned}\mathbf{h}(\mathbf{u}^R, \mathbf{u}^L; \mathbf{n}) &= \frac{1}{2}\left(\mathbf{f}(\mathbf{u}^R, \mathbf{n}) + \mathbf{f}(\mathbf{u}^L; \mathbf{n})\right) \\ &- \frac{1}{2}|A(\mathbf{u}^R, \mathbf{u}^L; \mathbf{n})|\left(\mathbf{u}^R - \mathbf{u}^L\right) \qquad (3.6)\end{aligned}$$

   where $\mathbf{f}(\mathbf{u}; \mathbf{n}) = \mathbf{F}(\mathbf{u}) \cdot \mathbf{n}$, and $A = d\mathbf{f}/d\mathbf{u}$ is the flux Jacobian.

2. *Componentwise limiting.* The solution variables are reconstructed componentwise. In principle, any set of variables can be used in the reconstruction (primitive variables, entropy variables, etc.). Note that conservation of the mean can make certain variable combinations more difficult to implement than others because of the nonlinearities that may be introduced. The simplest choice is obviously the conserved variables themselves. When conservation of the mean is not important (steady-state calculations), we typically use primitive variables in the reconstruction step.

3. *Weak Boundary Conditions.* Boundary conditions for inviscid flow at solid surfaces are enforced weakly. For solid wall boundary edges, the flux is calculated with $\mathbf{V} \cdot \mathbf{n}$ set identically to zero

$$\mathbf{f}(\mathbf{u}; \mathbf{n}) = \begin{pmatrix} 0 \\ n_x p \\ n_y p \\ 0 \end{pmatrix}.$$

   Boundary conditions at far field boundaries are also done weakly. Define the characteristic projectors of the flux Jacobian A in the following way:

$$P^{\pm} = \frac{1}{2}[I \pm \text{sign}(A)].$$

   At far field boundary edges the fluxes are assumed to be of the form:

$$\mathbf{f}(\mathbf{u}^n; \mathbf{n}) = (\mathbf{F}(\mathbf{u}^n_{proj}) \cdot \mathbf{n})$$

where $\mathbf{u}^n_{proj} = P^+ \ \mathbf{u}^n + P^- \ \mathbf{u}_\infty$ and $\mathbf{u}_\infty$ represents a vector of prescribed far field solution values. At first glance, prescribing the entire vector $\mathbf{u}_\infty$ is an overspecification of boundary conditions. Fortunately the characteristic projectors remove or ignore certain combinations of data so that the correct number of conditions are specified at inflow and outflow.

3.2.2 Example: Supersonic Oblique Shock Reflections

In this example, two supersonic streams (M=2.50 and M=2.31) are introduced at the left boundary. These streams interact producing a pattern of supersonic shock reflections down the length of the converging channel. The grid is a subdivided 15x52 mesh with perturbed coordinates as shown in Fig. 3.8. This same figure also shows Mach contours for the numerical solution obtained using piecewise constant reconstruction. As expected, the piecewise constant reconstruction scheme severely smears the shock system while the scheme based on a linear solution reconstruction shown in Fig. 3.9 performs very well. The piecewise quadratic approximation, shows some improvement in shock wave thickness although the improvement is not dramatic given the increased number of unknowns used in this particular scheme [Bar93]. The number of unknowns required for the quadratic approximation is roughly four times the number required for the piecewise linear scheme. This lack of dramatic improvement is not a surprising result since the solution has large regions of constant flow which do not benefit greatly from the quadratic approximation. At solution discontinuities the quadratic scheme reduces to a low order approximation which again negates the benefit of the quadratic reconstruction. To provide a fair comparison, Fig. 3.10 shows the same mesh adaptively refined. The number of mesh points has roughly doubled. A numerical solution was then obtained using linear reconstruction. The results are very comparable to the calculation performed using quadratic reconstruction while requiring less that 10% as much computational time.

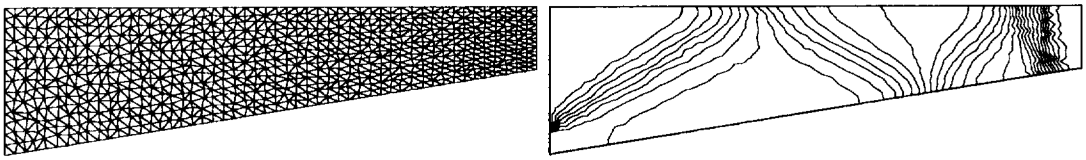

Figure 3.8: (a) Channel grid, 780 vertices (left). (b) Mach contours, piecewise constant reconstruction (right).

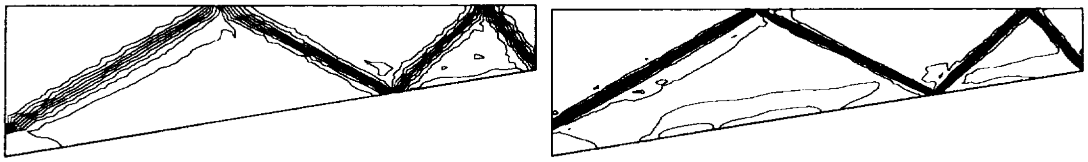

Figure 3.9: Mach contours. (a) Piecewise linear reconstruction (left). (b) Piecewise quadratic reconstruction (right).

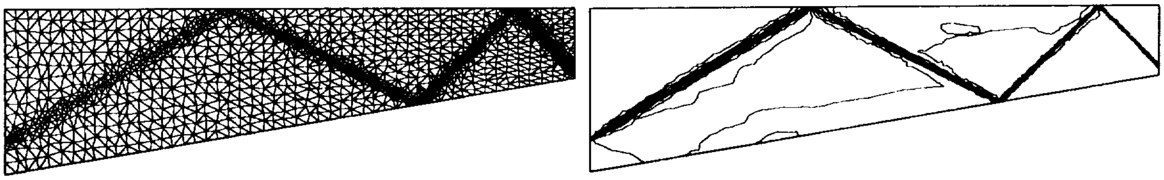

Figure 3.10: (a) Adapted channel grid, 1675 vertices. (left). (b) Mach contours, piecewise linear reconstruction (right).

3.2.3 Example: Transonic Airfoil Flow

Figure 3.11 shows a simple Steiner triangulation and the resulting solution obtained with a linear reconstruction scheme for transonic Euler flow ($M_\infty = .80, \alpha = 1.25°$) over a NACA 0012 airfoil section.

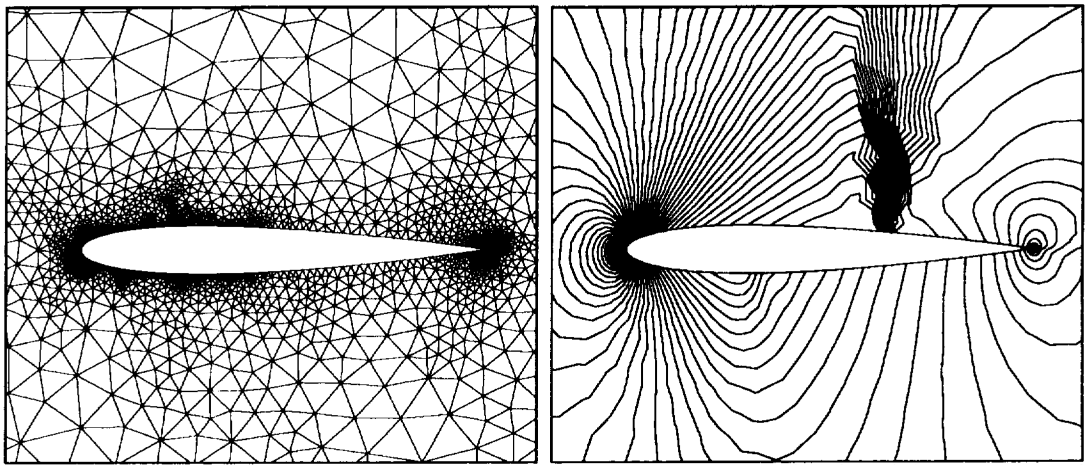

Figure 3.11: (a) Initial triangulation of airfoil, 3155 vertices (left). (b) Mach number contours, $M_\infty = .80, \alpha = 1.25°$.

Even though the grid is very coarse with only 3155 vertices, the upper surface shock is captured cleanly with a profile that extends over two cells of the mesh. Clearly, the power of the unstructured grid method is the ability to locally adapt the mesh to resolve flow features. Figure 3.12 shows an adaptively refined mesh and solution for the same flow. The flow features in Fig. 3.12 are clearly defined with a weak lower surface shock now visible. Figure 3.13 shows the surface pressure coefficient distribution on the airfoil. The discontinuities are monotonically captured by the scheme.

3.2.4 Example: Navier-Stokes Flow with Turbulence

As a last finite volume example, compressible Navier-Stokes flow is computed about a multiple-component airfoil geometry. In addition to the basic Navier-Stokes equations, the effects of turbulence on the mean flow equations are modeled using an eddy viscosity turbulence transport model. In a report with Baldwin [BB90], we proposed a single equation turbulence transport model with the specific application to unstructured meshes in mind.

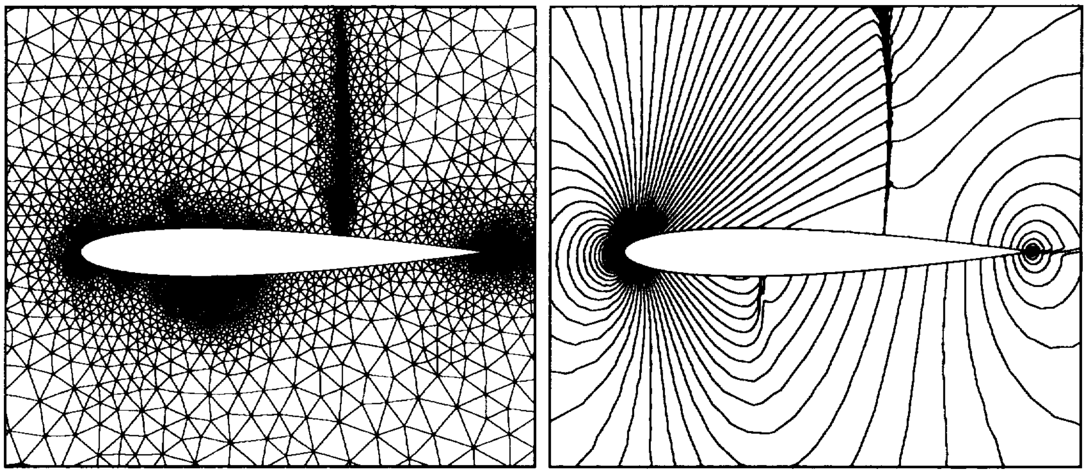

Figure 3.12: (a) Solution adaptive triangulation of airfoil, 6917 vertices. (b) Mach number solution contours on adapted airfoil.

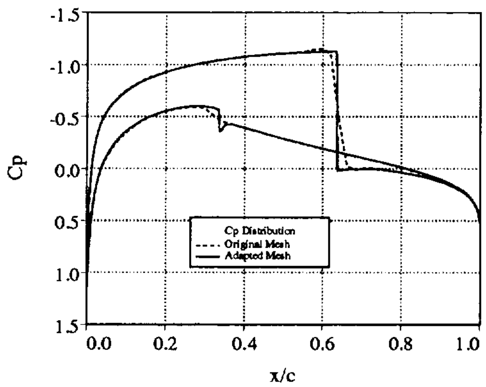

Figure 3.13: Comparison of C_p distributions on initial and adapted meshes.

This model was subsequently modified by Spalart and Allmaras [SA92] to improve the predictive capability of the model for wakes and shear-layers as well as to simplify the model's dependence on distance to solid walls. In the present computations, the Spalart model is solved in a form fully coupled to the Navier-Stokes equations. The one-equation model for the viscosity-like parameter $\tilde{\nu}$ is written

$$\frac{D\tilde{\nu}}{Dt} = \frac{1}{\sigma}\left[\nabla\cdot((\nu+\tilde{\nu})\nabla\tilde{\nu}) + c_{b2}(\nabla\tilde{\nu})^2\right] - c_{w1}f_w\left(\frac{\tilde{\nu}}{d}\right)^2 + c_{b1}\tilde{S}\tilde{\nu}. \tag{3.7}$$

In the Spalart model the kinematic eddy viscosity is given by $\nu_t = \tilde{\nu} f_{v1}$ and requires the following closure functions and constants

$$\tilde{S} = |\omega| + \frac{\nu\tilde{\nu}}{\kappa^2 d^2} f_{v2}, \quad f_{v1} = \frac{\chi^3}{\chi^3 + c_{v1}^3}$$

$$
\begin{aligned}
f_{v2} &= 1 - \frac{\chi}{1+\chi f_{v1}}, \quad r = \frac{\tilde{\nu}}{\tilde{S}\kappa^2 d^2} \\
g &= r + c_{w2}(r^6 - r)
\end{aligned}
$$

with ω the fluid vorticity, d the distance to the closest surface, and the constants $c_{b1} = 0.1355$, $c_{b2} = 0.622$, $c_{v1} = 7.1$, $c_{w1} = 3.24$, $c_{w2} = 0.3$, $c_{w3} = 2.0$, $\kappa = .41$, $\sigma = 2./3.$. The model also includes an optional term for simulating transition to turbulence. Using this model, viscous flow with turbulence is computed about the multiple-element airfoil geometry. This geometry has been triangulated using the Steiner triangulation algorithm described in [Bar95a], see Figure 3.14. The relatively coarse mesh contains approximately 22,000 vertices with cells near the airfoil surface attaining aspect ratios greater than 1000:1.

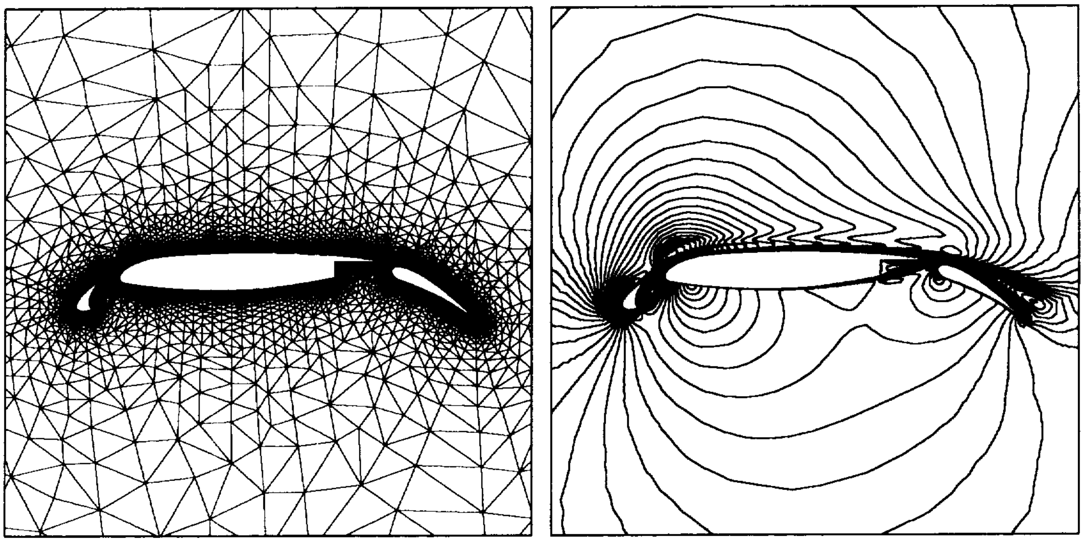

Figure 3.14: Multi-element airfoil mesh (left) and solution isomach contours (right), $M_\infty = 0.2$, $\alpha = 16.0°$, $Re = 9.0$ million.

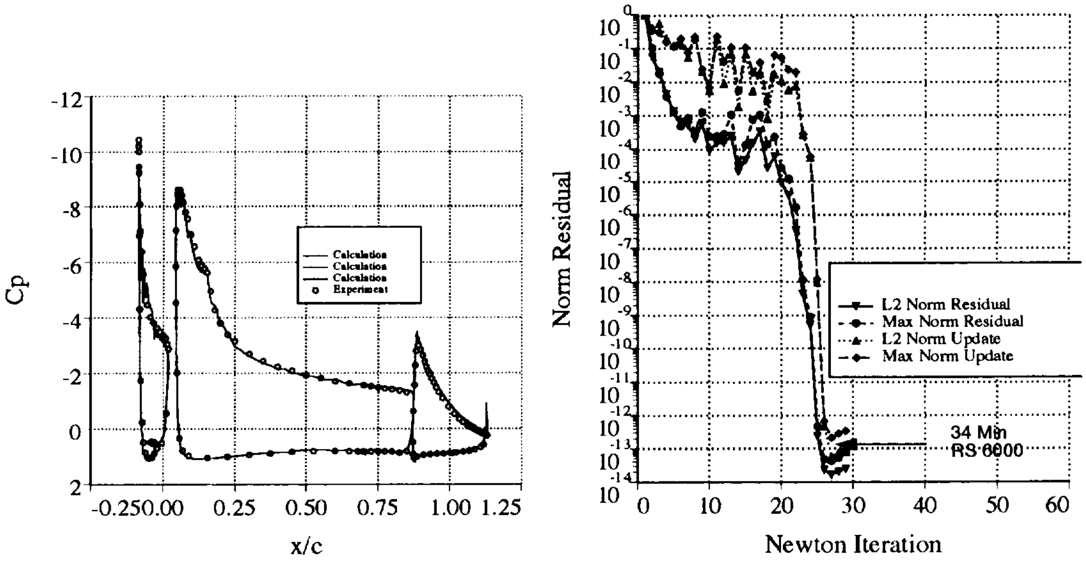

Figure 3.15: Comparison of computation and experiment (left), solution convergence history (right)

This example provides a demanding test case for CFD algorithms. The experimental flow conditions are $M_\infty = .20$, $\alpha = 16^\circ$, and a Reynolds number of 9 million. Experimental results are given in [VDMG92]. Even though the wake passing over the main element is not well resolved, the surface pressure coefficient shown in Figure 3.15 agrees quite well with experiment. The spatially discretized flow equations are solved using an exact Newton interation as described in [Bar95b]. The eventual quadratic convergence of Newton's method is shown in Fig. 3.15.

Chapter 4

Simplified GLS in Symmetric Form

This chapter returns to the symmetrization topic of Chapter 1 and presents a new simplified variant of the Galerkin least-squares (GLS) scheme in symmetric form. By combining the Scaling Theorem 1.2.1 of Chapter 1 with a simple congruence approximation given in the next section, a substancial simplification of this scheme is possible.

4.1 Congruence Approximation

Lemma 4.1.1 (Congruence Approximation) Let $A \in \mathbf{R}^{n\times n}$ be an arbitrary diagonalizable matrix and $B \in \mathbf{R}^{n\times n}$ a right symmetrizer of A. If $R \in \mathbf{R}^{n\times n}$ denotes the right eigenvector matrix scaled under Theorem 1.2.1 such that

$$A = R\Lambda R^{-1}, \quad \text{and } B = R\,R^T,$$

then the following inertially equivalent approximate decomposition exists

$$(A\,B)^{\pm} \approx R\,\Lambda^{\pm}\,R^T$$

satisfying

$$A\,B = (A\,B)^{+} + (A\,B)^{-} = R\Lambda^{+}R^T + R\,\Lambda^{-}\,R^T.$$

Proof: Let $U \in \mathbf{R}^{n\times n}$ denote the unitary matrix diagonalizing $A\,B$

$$A\,B = U\Omega U^T = R\Lambda R^T, \quad UU^T = I, \;\; RR^T = B.$$

Left and right multiplication by U^T and U respectively yields

$$\Omega = (U^T R)\Lambda(U^T R)^T.$$

From Sylvestor's law of inertia

$$Inertia(\Omega) = Inertia(\Lambda)$$

so consequently

$$Inertia(\Omega^{\pm}) = Inertia(\Lambda^{\pm}) = Inertia\left((U^T R)\Lambda^{\pm}(U^T R)^T\right).$$

Left and right multiplication by U and U^T respectively

$$U\Omega^{\pm}U^T = (A\,B)^{\pm} \cong R\Lambda^{\pm}R^T.$$

Finally, adding positive and negative components reveals

$$R(\Lambda^+ + \Lambda^-)R^T = R\Lambda R^T = A\,B.$$

■

4.2 Simplified Galerkin Least-Squares in Symmetric Form

Let $I^n =]t^n, t^{n+1}[$ denote the nth time interval and Ω the spatial domain composed of nonoverlapping elements T_i, $\Omega = \cup T_i$, $T_i \cap T_j = \emptyset$, $i \neq j$. Next define the trial and weighting finite element spaces

$$\mathcal{S}^h = \left\{\mathbf{v}^h | \mathbf{v}^h \in \left(C^0(\Omega \times I^n)\right)^m, \mathbf{v}^h|_{T\times I^n} \in \left(\mathcal{P}_k(T \times I^n)\right)^m, \mathbf{q}(\mathbf{v}) = \mathbf{g}_D \text{ on } \Gamma_D^- \times I^n\right\}$$

$$\mathcal{V}^h = \left\{\mathbf{w}^h | \mathbf{w}^h \in \left(C^0(\Omega \times I^n)\right)^m, \mathbf{w}^h|_{T\times I^n} \in \left(\mathcal{P}_k(T \times I^n)\right)^m, \mathbf{q}'(\mathbf{w}) = 0 \text{ on } \Gamma \times I^n\right\}$$

where $\mathbf{v}$ denotes the entropy variables for the system. The Galerkin least-squares method is defined by the following bilinear forms [HFM86]:

Find $\mathbf{v}^h \in \mathcal{V}^h$ such that for all $\mathbf{w}^h \in \mathcal{S}^h$

$$B(\mathbf{v}^h, \mathbf{w}^h)_{gal} + B(\mathbf{v}^h, \mathbf{w}^h)_{ls} + B(\mathbf{v}^h, \mathbf{w}^h)_{bc} = 0$$

with

$$\begin{aligned}
B(\mathbf{v}, \mathbf{w})_{gal} &= \int_{I^n}\int_{\Omega}(-\mathbf{u}(\mathbf{v}) \cdot \mathbf{w}_t - \mathbf{f}^i(\mathbf{v})\mathbf{w}_{,x_i}\, d\Omega\, dt \\
&+ \int_{\Omega}(\mathbf{w}(t_-^{n+1})\mathbf{u}(\mathbf{v}(t_-^{n+1})) - \mathbf{w}(t_+^n)\mathbf{u}(\mathbf{v}(t_-^n)))\; d\,\Omega \\
B(\mathbf{v}, \mathbf{w})_{ls} &= \int_{I^n}\sum_{T\in\Omega}\int_T \left(A^0\mathbf{w}_t + A^iA^0\mathbf{w}_{,x_i}\right)\boldsymbol{\tau}\left(A^0\mathbf{v}_t + A^iA^0\mathbf{v}_{,x_i}\right)\, d\Omega\, dt \\
B(\mathbf{v}, \mathbf{w})_{bc} &= \int_{I^n}\int_{\Gamma_F} w\;\mathbf{h}(\mathbf{v}, \mathbf{g}_F; \mathbf{n})\; d\Gamma\; dt
\end{aligned}$$

where

$$\mathbf{h}(\mathbf{v}_-, \mathbf{v}_+, \mathbf{n}) = \frac{1}{2}(\mathbf{f}(\mathbf{u}(\mathbf{v}_-); \mathbf{n}) + \mathbf{f}(\mathbf{u}(\mathbf{v}_+); \mathbf{n})) - \frac{1}{2}|A(\mathbf{u}(\overline{\mathbf{v}}); \mathbf{n})|(\mathbf{u}(\mathbf{v}_+) - \mathbf{u}(\mathbf{v}_-)).$$

4.2.1 A Simplified Least-Squares Operator in Symmetric Form

In the implementation of the Galerkin least-squares method, the most difficult computational aspect of the scheme is the calculation of the least-squares $\boldsymbol{\tau}$ matrix. In the paper by Hughes and Mallet [HM86], they proposed the following general form for $\boldsymbol{\tau}$ on a mapped, $\mathbf{x} \mapsto \boldsymbol{\xi}$, anisotropic element

$$\boldsymbol{\tau}_p = |\tilde{B}|_p^{-1}, \quad |\tilde{B}|_p = \left(\sum_{i=0}^{d}|\tilde{B}^i|^p\right)^{1/p}, \tilde{B}^i = (\partial\xi_i/\partial x_j)A^jA^0 \tag{4.1}$$

where A^j and A^0 are the matrices given in (1.5). Historically, the value $p = 2$ has been used in implementations of the Galerkin least-squares method. After manipulation, τ_2 can be written in the form

$$\tau_2 = (A^0)^{-1} \left(\left(\frac{\partial \xi_0}{\partial x_0} \right)^2 + \left(\frac{\partial \xi_i}{\partial x_j} \right) \left(\frac{\partial \xi_i}{\partial x_k} \right) A^j A^k \right)^{-1/2} .$$

This necessitates the calculation of a matrix square root. Hughes advocates the use of the Cayley-Hamilton theorem for this computation. Unfortunately, the Cayley-Hamilton technique becomes unwieldy for matrices of dimension larger than 4 or 5. In light of the Scaling Theorem 1.2.1, it is useful to revisit the derivation of τ with $p = 1$. Define $\mathbf{n}_i = \nabla\xi_i / |\nabla\xi_i|$ and $A(\mathbf{n}) = n_i A^i$ so that (4.1) can be rewritten as

$$\tau_1 = |\tilde{B}|_1^{-1} = \Big(|\nabla\xi_i| \, |A(\mathbf{n}_i) A^0| \Big)^{-1} .$$

Observe that for the Euler and MHD equations considered in Chapter 1, the eigenvectors of $A(\mathbf{n})A^0$ are not easily obtained. Hence, the computation of $|A(\mathbf{n}_i)A^0|$ seems nontrivial as well. Lemma 4.1.1 suggests the inertially equivalent approximation

$$|A(\mathbf{n})A^0| \approx R(\mathbf{n}) |\Lambda(\mathbf{n})| R^T(\mathbf{n})$$

using the entropy scaled eigenvectors $R(\mathbf{n})$ of $A(\mathbf{n})$. From this, the following approximation for τ follows

$$\tau_1 \approx \Big(|\nabla\xi_i| \, R(\mathbf{n}_i) |\Lambda(\mathbf{n}_i| R^T(\mathbf{n}_i) \Big)^{-1} .$$

This represents a substancial simplification of the τ matrix calculation.

4.2.2 Example: MHD Flow for Perturbed Prandtl-Meyer Fan

In this example, the MHD equations are solved in the square domain $\Omega \in [-1/2, 1/2]^2$ using the simplified GLS scheme. The inflow data consists of a Mach 1.55 Prandtl-Meyer fan with origin located at ($x = -.881, y = .47147$). A velocity aligned magnetic field, $\mathbf{B} = .05\mathbf{V}$, is introduced at the boundary. Figure 4.1 shows the coarsest mesh used for the computation and Mach number contours computed from the simplified GLS scheme. Figure 4.2 shows contours of the x-component of the magnetic field using linear and quadratic elements on the coarsest mesh. Solutions were then obtained on a sequence of meshes with decreasing mesh spacing and the L_2 norm of $\nabla \cdot \mathbf{B}$ measured. Figure 4.3 graphs the convergence of $\nabla \cdot \mathbf{B}$ for linear and quadratic approximations. The graphs show optimal convergence rates, $O(h)$ for linear and $O(h^2)$ for quadratic elements.

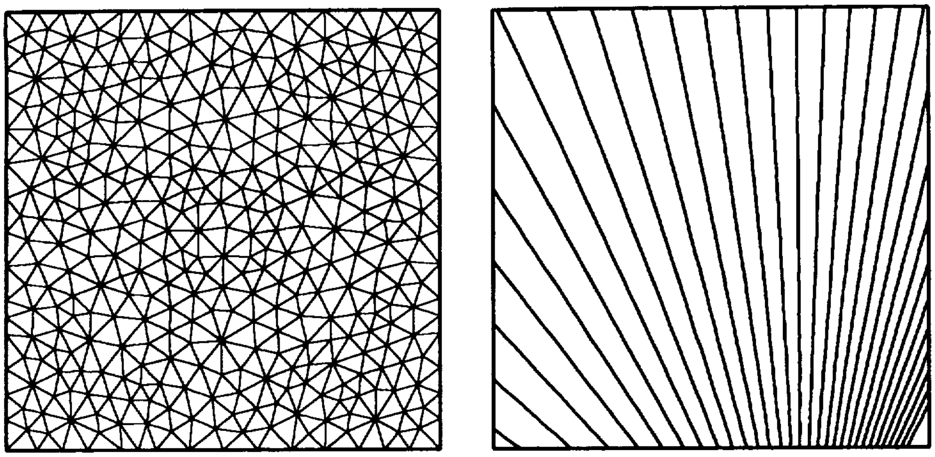

Figure 4.1: Coarsest mesh (400 vertices) and Mach number contours for perturbed Prandtl-Meyer fan.

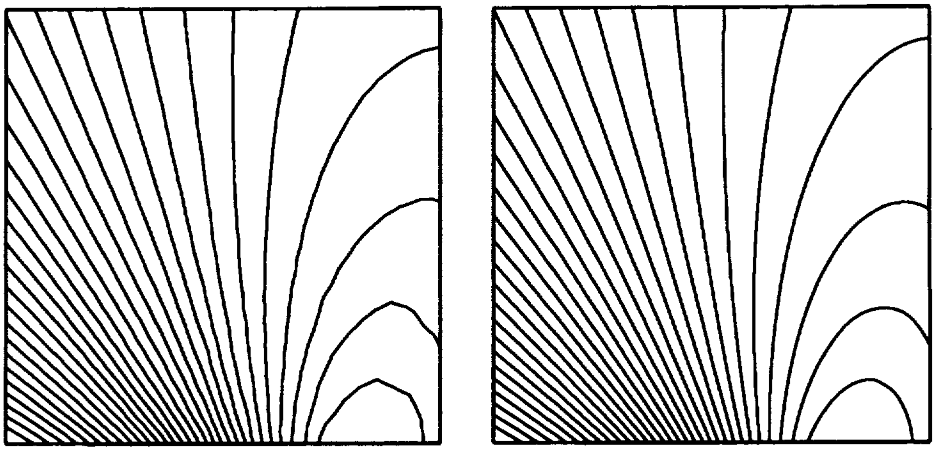

Figure 4.2: B_x component of magnetic field computed on coarsest mesh. Galerkin least-squares with linear elements (left) and quadratic elements (right).

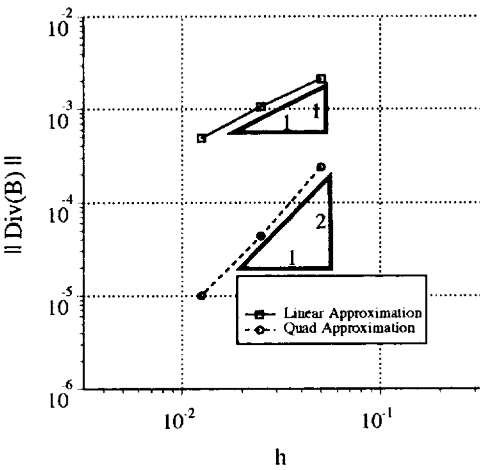

Figure 4.3: Spatial Convergence of $\nabla \cdot \mathbf{B}$ for the Galerkin least-squares scheme using linear and quadratic elements.

Bibliography

[Abg94] R. Abgrall. An essentially non-oscillatory reconstruction procedure on finite-element type meshes. *Comp. Meth. Appl. Mech. Eng.*, 116:95–101, 1994.

[Abg95] R. Abgrall. On essentially non-oscillatory schemes on unstructured meshes. *J. Comp. Phys.*, 114:45–58, 1995.

[AD93] P. Arminjon and A. Dervieux. Construction of tvd-like artificial viscosities on two-dimensional arbitrary fem grids. *J. Comp. Phys.*, 106(1):176–198, 1993.

[Bal94] D. Balsara. Higher-order Godunov schemes for isothermal hydrodynamics. *Astrophysical J.*, 420:197–203, 1994.

[Bar91] T. J. Barth. Unstructured grids and finite-volume solvers for the Euler and Navier-Stokes equations, March 1991. von Karman Institute Lecture Series 1991-05.

[Bar93] T. J. Barth. Recent developments in high order k-exact reconstruction on unstructured meshes. Technical Report AIAA 93-0668, Reno, NV, 1993.

[Bar94] T. J. Barth. Aspects of unstructured grids and finite-volume solvers for the Euler and Navier-Stokes equations, March 1994. von Karman Institute Lecture Series 1994-05.

[Bar95a] T. J. Barth. Steiner triangulation for isotropic and stretched elements. Technical Report AIAA 95-0213, Reno, NV, 1995.

[Bar95b] T.J. Barth. Parallel cfd algorithms on unstructured meshes. Technical Report AGARD Report R-907, Advisory Group for Aerospace Research and Development, 1995.

[BB90] B. S. Baldwin and T. J. Barth. A one-equation turbulence transport model for high Reynolds number wall-bounded flows. Technical Report TM-102847, NASA Ames Research Center, Moffett Field, CA, August 1990.

[Bey91] K. Bey. A Runge-Kutta discontinuous finite element method for high speed flows. Technical Report AIAA 91-1575, American Institute for Aeronautics and Astronautics, Honolulu, Hawaii, 1991.

[BJ89] T. J. Barth and D. C. Jespersen. The design and application of upwind schemes on unstructured meshes. Technical Report AIAA 89-0366, American Institute for Aeronautics and Astronautics, Reno, NV, 1989.

[BW88] M. Brio and C.C. Wu. An upwind differencing scheme for the equations of ideal magnetohydrodynamics. *J. Comp. Phys.*, 75:400–422, 1988.

[CM80] M.G. Crandall and A. Majda. Monotone difference approximations for scalar conservation laws. *Math. Comp.*, 34:1–21, 1980.

[CR73] P.G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comp. Meth. in Appl. Mech. and Eng.*, 2:17–31, 1973.

[CW84] P. Collela and P. Woodward. The piecewise parabolic methods for gas-dynamical simulations. *J. Comp. Phys.*, 54, 1984.

[DD88] J. Desideri and A. Dervieux. Compressible flow solvers using unstructured grids, March 1988. von Karman Institute Lecture Series 1988-05.

[Del34] B. Delaunay. Sur la sphére vide. *Izvestia Akademii Nauk SSSR*, 7(6):793–800, 1934.

[Gan59] F.R. Gantmacher. *Matrix Theory.* Chelsea Publishing Company, New York, N.Y., 1959.

[God59] S.K. Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Mat. Sb.*, 47, 1959.

[God61] S.K. Godunov. An interesting class of quasilinear systems. *Dokl. Akad. Nauk. SSSR*, 139:521–523, 1961.

[GV85] J.D. Goodman and R.J. Le Veque. On the accuracy of stable schemes for 2d conservation laws. *Math. Comp.*, 45, 1985.

[Har83a] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comp. Phys.*, 49:357–393, 1983.

[Har83b] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comp. Phys.*, 49:151–164, 1983.

[HFM86] T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for CFD: I. symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comput. Meth. Appl. Mech. Eng.*, 54:223–234, 1986.

[HHL76] A. Harten, J. M. Hyman, and P. D. Lax. On finite-difference approximations and entropy conditions for shocks. *Comm. Pure and Appl. Math.*, 29:297–322, 1976.

[HM86] T.J.R. Hughes and M. Mallet. A new finite element formulation for CFD: III. the generalized streamline operator for multidimensional advective-diffusive systems. *Comput. Meth. Appl. Mech. Eng.*, 58:305–328, 1986.

[HOEC87] A. Harten, S. Osher, B. Enquist, and S. Chakravarthy. Uniformly high-order accurate essentially nonoscillatory schemes III. *J. Comp. Phys.*, 71(2):231–303, 1987.

[JL86] A. Jameson and P. Lax. Conditions for the construction of multi-point total variation diminishing difference schemes. Technical Report ICASE report 178076, NASA Langley, 1986.

[JS90] C. Johnson and A. Szepessy. Convergence of the shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws. *Math. Comp.*, 54:107–129, 1990.

[KRW96] D. Kröner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-d. *East-West J. Numer. Math.*, 4(4):279–292, 1996.

[KSR95] D. Kröner, S. Sebastian, and M. Rokyta. Convergence of higher order upwind finite volume schemes on unstructured grids for scalar conservation laws in several space dimensions. *Numer. Math.*, 71(4):527–560, 1995.

[Law86] C. Lawson. Properties of n-dimensional triangulations. *CAGD*, 3:231–246, 1986.

[Lax73] P.D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves.* SIAM, Philadelphia, Penn., 1973.

[Mic94] C.R. Michel. Improved reconstruction schemes for the navier-stokes equations on unstructured meshes. Technical Report AIAA 94-0642, American Institute for Aeronautics and Astronautics, 1994.

[Moc80] M. S. Mock. Systems of conservation laws of mixed type. *J. of Differ. Equations*, 37:70–88, 1980.

[Pow94] K. G. Powell. An approximate Riemann solver for magnetohydrodynamics (that works in more than one dimension). Technical Report ICASE 94-24, NASA Langley, 1994.

[RB96] P. L. Roe and D.S. Balsara. Notes on the eigensystem of magnetohydrodyamics. *SIAM J. Appl. Math.*, 56(1):57–67, 1996.

[Roe81] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43, 1981.

[RS81] T. Ruggeri and A. Strumia. Convex convariant entropy density, symmetric conservative form, and shock waves in relativistic magnetohydrodynamics. *J. Math. Phys.*, 22(8):1824–1827, 1981.

[SA92] P. Spalart and S. Allmaras. A one-equation turbulence model for aerodynamic flows. Technical Report AIAA 92-0439, Reno, NV, 1992.

[Spe87] S. Spekreijse. *Multigrid Solution of the Steady Euler-Equations.* PhD thesis, Centrum voor Wiskunde en Informatica, Amsterdam, 1987.

[Str89] R. Struijs. An adaptive grid polygonal finite volume method for the compressible flow equations. Technical Report AIAA 89-1959-CP, American Institute for Aeronautics and Astronautics, 1989.

[Str94] R. Struijs. *A Multi-Dimensional Upwind Discretization Method for the Euler Equations on Unstructured Grids.* PhD thesis, T.U. Delf and the VKI Institute, 1994.

[Van93] P. Vankeirsbilck. *Algorithmic Developments for the Solution of Hyperbolic Conservation Laws on Adaptive Unstructured Grids.* PhD thesis, Katholiek Universiteit van Leuven, 1993.

[VDMG92] W. Valarezo, C. Dominik, R. McGhee, and W. Goodman. *High Reynolds Number Confguration Development of a High-Lift Airfoil.* Technical Report AGARD Meeting In High-Left Aerodynamics 10-01, 1992.

[Ven93] V. Venkatakrishnan. On the accuracy of limiters and convergence to steady state. Technical Report AIAA 93-0880, Reno, NV, 1993.

[vL79] B. van Leer. Towards the ultimate conservative difference schemes v. a second order sequel to Godunov's method. *J. Comput. Phys.*, 32, 1979.